JMIR Infodemiology

Journal Impact Factor (JIF) (2023): 3.5
Volume 6 (2026) ISSN 2564-1891 Editor-in-Chief: Tim Ken Mackey, MAS, PhD

Contents

Original Papers

Marketing Strategies and Factors Influencing the Popularity of Alcohol Videos from Official Brand Accounts on Douyin: Content Analysis Study

Yuchen Zhao[1], MB; Lingyun Zhang[1,2], MPH; Chenyu Qian[1], MPH; Wenjie Guo[1], MB; Weiyun Zhu[1], MB; Pinpin Zheng[1], PhD

[1]Department of Preventive Medicine and Health Education, School of Public Health, Institute of Health Communication, Fudan University, 138 Yixueyuan Road, Xuhui District, Shanghai, China

[2]Party Committee Office, Zhongshan Hospital Affiliated with Fudan University, Shanghai, China

Corresponding Author:
Pinpin Zheng, PhD
Department of Preventive Medicine and Health Education, School of Public Health, Institute of Health Communication, Fudan University, 138 Yixueyuan Road, Xuhui District, Shanghai, China

Abstract

Background: Alcohol consumption in China poses significant public health challenges. Alcohol marketing has been shown to increase public alcohol consumption, with social media platforms such as Douyin (TikTok in Mainland China) being among the main channels for alcohol marketing.

Objective: This study aimed to analyze the thematic content of alcohol advertising on the Douyin platform and to explore the factors influencing the popularity of these types of advertising.

Methods: Using data from the JINGDONG platform and alcohol industry reports, we identified 40 popular alcohol brands. For each brand, we located their official Douyin accounts and selected the top 20 most-liked videos posted between November 1, 2020, and November 1, 2021. In total, 659 videos from 37 brands were collected for analysis. Two trained researchers independently coded each video using a predefined codebook, which consisted of 7 sections and 20 items. Binary logistic regression was conducted with the grouping of the number of likes as the dependent variable, and the marketing strategies and warning elements of each video as independent variables.

Results: Among the 659 videos analyzed, 320 (48.6%) garnered more than 1000 likes. A significant portion of the videos was direct advertisements (281/659, 42.6%) and short skits (255/659, 38.7%), with 56.0% (369/659) featuring characters engaging in drinking-related behaviors or directly consuming alcohol. Additionally, many videos highlighted brand elements (510/659, 77.4%) and extended features (161/659, 24.4%). Cultural themes were also common, with 23.2% (153/659) of the videos promoting the enjoyment of life and 6.8% (45/659) emphasizing balance in life. However, age restrictions were missing for 26.9% (177/659) of the videos, and only 1.2% (8/659) included a health warning stating that “Drinking is harmful to health.” Certain marketing strategies were significantly associated with greater video popularity, including the use of short skits (odds ratio [OR] 2.77, 95% CI 1.42 - 5.41), highlighting brand elements (OR 2.96, 95% CI 1.59 - 5.51), and emphasizing life balance (OR 3.44, 95% CI 1.11 - 10.66). In contrast, the presence of age restrictions (OR 0.32, 95% CI 0.15 - 0.67) and explicit health warnings (OR 0.06, 95% CI 0.01 - 0.84) were associated with lower popularity. The period from July to September and November was the peak release period for alcohol advertisements on Douyin.

Conclusions: Alcohol marketing strategies on Douyin leverage experiential, brand-driven, collaborative, and cultural marketing techniques to enhance video attractiveness and create alcogenic environments. Moreover, effective age restrictions and health warnings are largely absent. It is essential to legislate and enforce stricter alcohol marketing regulations to reduce the health risks associated with alcohol marketing.

(JMIR Infodemiology 2026;6:e74221) doi:10.2196/74221

KEYWORDS

alcohol marketing; Douyin platform; alcogenic environments; advertisement; health warning

Introduction

Alcohol consumption is a leading risk factor for the global disease burden, which is associated with the risk of more than 200 diseases and injuries, causing approximately 3 million deaths globally each year, accounting for 5.3% of all deaths [1]. There is growing evidence that the public's alcohol consumption could be influenced by "alcogenic environments," which are settings in which alcohol is easily accepted, available, and affordable [2]. "Alcogenic environments" have been shown to be associated with harmful drinking behaviors, including alcohol addiction and underage drinking [3-5].

As an important part of the "commercial determinants of health," alcohol marketing is a key driver for creating an alcogenic environment [2,5]. The alcohol industry often promotes products through intensive advertising and even emphasizes the "function" of drinking [3]. This marketing strategy of widespread advertising could help enhance the acceptability and normalization of alcohol consumption at both the individual and community levels, resulting in an increase in the alcogenic environment [5-8].

The widespread use of digital media has provided new channels for alcohol marketing, such as websites, apps, and social media [9,10]. Alcohol marketing on social media platforms has grown rapidly over the past two decades, with the alcohol industry reporting that alcohol marketing on social media platforms (including YouTube and Facebook) has reached many more consumers [11]. For instance, a previous study revealed that music videos on YouTube conveyed approximately 10.06 billion alcohol-related impressions to the British population [12].

As a country with high alcohol consumption, alcohol marketing on social media in China is also severe [13]. The annual advertising investment of the Chinese liquor industry has exceeded 20 billion yuan (approximately US $3 billion), with the goal of creating an all-media marketing platform combining television, websites, and social media [14]. A Chinese alcohol industry report in 2021 informed that 89% of alcohol consumers received alcohol-related information through the internet, with more than 30% coming from short video platforms such as Douyin (TikTok in mainland China) [15].

Douyin, China's leading short video platform with more than 800 million daily active users, is primarily youth-centric, with 60% of its users aged younger than 30 years [16]. It has become a major hub for entertainment, social interaction, and marketing, leveraging algorithm-based recommendations to help users discover tailored content and enabling businesses to reach a broad audience with creative campaigns. Under the "Chinese liquor" tag, videos related to Chinese liquor on Douyin have a total of 65.63 billion views, with the most popular video receiving as many as 878,000 likes [17]. These videos feature diverse drinking scenes and emotions, using promotional strategies, such as collaborations with opinion leaders and brand partnerships [18].

According to the attention-interest-desire-action (AIDA) model, the consumer purchase decision-making process is characterized by a progression through four stages [19]: attention, interest, desire, and action. Previous research has demonstrated that higher levels of alcohol social media marketing exposure are associated with positive drinking expectancies and drinking behaviors [20,21], confirming that the AIDA model can be used to explain alcohol advertising exposure, where attention attracted by alcohol advertisements can translate into drinking behavior. However, regarding the "A (attention)" and "I (interest)" components of the AIDA model, likes can be viewed as a preliminary quantitative indicator of successful "interest" capture, and it remains unclear from previous research what types of alcohol-related social media advertisements effectively capture audience attractiveness.

The World Health Organization (WHO) has called for interventions to reduce the alcogenic environment, and the key interventions include restrictions on alcohol marketing [2]. However, China lacks comprehensive regulations governing social media alcohol advertising. It is essential to determine the marketing strategies of alcohol products to formulate more comprehensive regulations and supervision measures.

This study aimed to identify the marketing strategies of alcohol advertisements and the placement of warnings on the Douyin platform, the factors associated with the attractiveness of those alcohol advertisements, and the time trend of these alcohol videos. These findings may contribute to extending the application of the AIDA model in social media alcohol advertising, with a specific focus on exploring the mechanisms through which different types of alcohol marketing strategies facilitate the transition from the "attention" to the "interest" stage. These findings are also expected to provide policymakers with insights into how alcohol products are promoted and marketed on social media, which is crucial for enhancing the regulation of alcohol marketing.

Methods

Sampling and Data Collection

In 2021, the size of the online market for alcohol in China reached 136.31 billion yuan [22], of which JINGDONG [23] was one of the main Chinese e-commerce platforms. The sales rankings of alcohol brands on JINGDONG can reflect the popularity of those brands. By reviewing alcohol industry reports, we categorized the alcohol currently sold in China into six groups: Chinese liquor (spirits), beer, wine, yellow rice wine, fruit wine, and premixed cocktails. On the basis of this classification, we determined the alcohol brands for this study as follows:

- Database A: According to retail data from the JINGDONG platform in 2021, the top 5 alcohol brands in each category were selected, resulting in a total of 30 brands representing popular choices across all age groups. These brands represent products commonly purchased by general consumers.
- Database B: In addition to the brands aforementioned, some new brands are very popular among young people and women but lack large sales records. We identified the 10 most popular brands targeting these demographics by

reviewing industry reports on alcohol consumption by young people and women.

- Combining databases A and B, a total of 40 brands were included in the study. We then identified the official Douyin accounts for each brand. Three brands did not have official Douyin accounts, leaving 37 brands for inclusion (Table 1). Given the similar formats of Douyin advertisements within the same brand and for the feasibility of the study, we selected the top 20 most-liked videos from each account that were released between November 1, 2020, and November 1, 2021. For brands that had released fewer than 20 videos in the past year, all available videos were included. The selection of 2021 as the data collection window was based on a key transition in Douyin's commercialization process: before 2020, Douyin's strategy focused primarily on user acquisition, while from 2020 onward, its commercialization accelerated, with the platform achieving the top position in domestic advertising revenue that same year [24]. This makes 2020 - 2021 a suitable period for examining the marketing strategies of alcohol brands during the platform's commercialization push. In total, 659 videos were collected for further analysis.

Table . Samples of alcohol brands in this study.

Categories	Brands
Chinese liquor	MOUTAI, WuLiangYe, Luzhou Laojiao, Yanghe, Fen Jiu, and Jiang Xiaobai
Beer	Budweiser, Tsingtao, Snowflake, Yanjing, Harbin, and Corona
Yellow or rice wine	Guyuelongshan, Kuaijishan, Jimo rice wine, Nverhong, MIK, Suzhou Qiao
Wine	Penfolds, Greatwall, Jacob's Creek, Lafite, Chateau Monlot, Changyu/Torre Oria
Fruit wine or preconditioning of cocktails	Jin liquor, Jinro, UMEET, Huatian Xiangz, RIO, Breezer, Power Station, Jiushilang
Imported liquor	Absolut Vodka, Bacardi, Johnnie Walker, Ballantine, and Hennessy

Coding Method

Codebook development was a top-down or bottom-up process. This process included the following phases. First is the top-down phase—reviewing the previous studies and the related advertising marketing reports [9,25,26], which allowed us to identify the marketing strategies used by official alcohol accounts on Douyin. Afterward, a preliminary framework of the codebook was developed with 8 main dimensions, including basic information, content presentation, scene setting, brand and product appeal, promotion strategy, emotion, culture, and warning (Multimedia Appendix 1). Second is the bottom-up phase—reviewing 20 randomly selected videos of official alcohol accounts on Douyin, which helped us identify the strategies that were undetermined in the preliminary framework. Accordingly, we enriched the codebook with subdimensions derived from the video content, including the specific form of the presentation, the characters' drinking action, the scene, the product promotion strategy, the emotional tone, the cultural appeal, and different types of warnings. Third, 20 videos were randomly selected from the remaining samples for retested coding to verify the applicability of the current codebook. The resulting codebook consisted of 8 sections with 19 items, including basic information, content presentation, scene setting, brand and product appeal, promotion strategy, emotion, culture, and warning (Table 2).

Apart from the aforementioned elements, this research also collected the like counts of each video as the variable of video attractiveness. As 1000 is close to the median number of likes (Median=870, IQR: 135-7612), we categorized the number of likes into a high likes group and a low likes group using 1000 as the cutoff point.

The content analysis of the alcohol-related video started in December 2021 and took 1 month in total. A total of 659 videos were monitored. A total of 3 public health researchers participated in the coding work, and all researchers received training to unify the coding standards before the start of coding. After the training, the consistency of the researchers when coding the same video reached a kappa value of 92.8%, indicating a high level of consistency in their coding standards.

The specific coding work was conducted via an online questionnaire platform "Wenjuanxing." The codebook was preentered into the platform and presented in the form of choice questions. Researchers selected the corresponding options on the online questionnaire platform while viewing Douyin videos. All videos were downloaded to local computers and coded independently by 2 researchers. If there were inconsistencies, the 2 researchers were required to rereview the video to confirm it. If they still failed to reach an agreement, the third researcher was involved in the discussion to reach a consensus.

Table . Coding book of alcohol videos on the Douyin platform.

Dimensions, variable, and category	Definition or example
Basic information	
Durations	
≤30	—[a]
31-60	—
≥61	—
Brand category	
Traditional	Brands from database A
New	Brands from database B
Content presentation	
Form of presentation	
Advertisement	A direct promotional message aimed at selling a product
Short skit	A brief video performance using theatrical elements
Vlog	Short-form documentaries
Film and TV[b] show excerpts	The clips from film and TV shows
Other	Other forms
Characters' drinking action	
No characters and no drinking behavior	No characters appear
Characters appear without drinking behavior	Characters do not touch alcohol products
Characters appear with drinking-related behavior	Characters hold alcohol glasses or pour alcohol and clink glasses
Characters appear with drinking behavior directly	Characters drink alcohol directly
Scene setting	
Drinking alone	Contain scenes of drinking alone
Party	Contain scenes of drinking at parties
Natural scenery	Contain scenes of natural scenery
Cultural or sports activities	Contain scenes of cultural or sports activities
Brand and product appeal	
Brand elements	Contain brand name, brand logo, or brand mascot
Product elements	Contain product name or product logo
Intrinsic product features	Contain information about the odor, color, taste, and materials of the product
Extended product features	Contain information about the origin, production process, vintage, and creative drinking methods
Promotion strategy	
Product promotion strategy	
Key opinion leaders	Invite celebrities to endorse
Cross-border brand cooperation	Collaborate with other brands to promote
Interaction with audience	Engage with the audience to attract fans
Not mentioned	—
Cues refer to women's interests	Contain cues that refer to women's interests, such as flowers, perfume
Cues refer to youth's interests	Contain cues that refer to youth interests, such as cartoons, cosplay
Emotion	
Emotional tone	

Dimensions, variable, and category	Definition or example
Positive	A favorable or optimistic emotion conveyed in the video, such as pleasure, moving
Neutral	An emotion neither leans positive nor negative, such as calmness
Negative	An unfavorable or unpleasant emotion conveyed in the video, such as sadness, loss
Culture	
Cultural appeal	
Historical inheritance	Highlighting the alcohol's history, heritage, and tradition
Festival celebration	Emphasizing drinking as a way to celebrate the holiday
Balance in life	Emphasize the philosophical concept of balance in life gained through drinking
Ambition and striving	Drinking symbolizes ambition and striving for success
Enjoy one's life	Emphasizing drinking as a way to enjoy life
Not mentioned	—
Warning	
Age restriction	
Yes or no	Contain age restrictions, such as "minors under the age of 18 are prohibited from drinking alcohol"
Health warnings	
Drinking is harmful to health	Contain health warning related to "drinking is harmful to health"
Please drink responsibly	Contain health warning related to "please drink responsibly"
Not mentioned	—

[a]Not applicable.

[b]TV: television.

Statistical Analysis

Frequencies and proportions are reported for the basic information, the marketing elements, and the warnings of Douyin videos. Chi-square tests were conducted to examine the association between different marketing strategies and the grouping of like counts, as well as the association between the warnings in videos and the grouping of like counts. As the dependent variable, "grouping of like counts," is a binary variable, we conducted a binary logistic regression, with the low-likes group serving as the reference. The marketing elements and warning elements of each video, which were recorded through coding, served as the independent variables. The enter method was adopted for the model to explore the factors affecting the popularity of Douyin videos among the public.

Adjusted odds ratios (ORs) and their 95% CIs were used to quantify the effects. To evaluate the model's goodness of fit, the Nagelkerke R^2 was used to determine the explanatory power of the model. To investigate the seasonal variations in Douyin videos, this study compiled the release times of the sampled videos and generated a scatter plot depicting the monthly distribution of video releases. IBM SPSS software (version 20.0) was used to carry out all the analyses.

Ethical Considerations

According to Article 32 of China's National Health Commission, Ministry of Education, and Ministry of Science and Technology Document No. 4 in 2023 "Notice on Issuing the Measures for Ethical Review of Human Life Science and Medical Research," research using legally obtained publicly available data, data generated through observation without interfering with public behavior, or anonymized information data is exempt from ethical review [27]. This study uses legally accessible public data from social media platforms, involves no individual user data, and does not interfere with public behavior. Therefore, this study qualifies for an exemption from ethical review.

Results

The Marketing Strategies of Alcohol Advertisements and the Placement of Warnings

Among the 659 Douyin videos analyzed, the number of likes ranged from 2 to 440,000, with a median of 870 (IQR: 135-7612). Using 1000 likes as the cutoff point, 320 videos (48.6%) were classified as high-liked videos, with more than 1000 likes, whereas 339 videos (51.4%) were classified as low-liked videos, with fewer than 1000 likes.

Most videos were presented as advertisements (n=281, 42.6%) and short skits (n=255, 38.7%), with 56.0% (n=369) of the characters engaging in drinking-related behavior or drinking

directly. The most frequent scenes were parties (n=170, 25.8%) and natural scenery (n=112, 17.0%). A total of 254 (38.5%) videos showed the product's intrinsic features, such as taste and odor, whereas 161 (24.4%) videos highlighted extended features, such as creative ways of mixing and drinking Rio cocktails with Sprite. Some videos conveyed drinking-related culture, mainly including enjoying one's life (n=153, 23.2%) and historical inheritance (n=65, 9.9%). For example, "Rainy days go better with blueberry wine" and "Tradition is our unwavering craftsmanship" (Figure 1).

Additionally, 36.6% (n=241) of the videos included elements favored by women, such as flowers and perfume, and 16.1% (n=106) contained elements appealing to teenagers, including e-sports and anime. However, not all videos included age restrictions (n=482, 73.1%), and only 1.2% (8/659) contained health warnings "Drinking is harmful to health" (Table 3).

Figure 1. Screens depicting the marketing strategy for alcohol advertisements on the Douyin platform. (A) Use of a short skit. (B) Creative method of drinking the product (adding grapefruit sauce to a cocktail). (C) Use of key opinion leaders. (D) The promotion of a life balance philosophy (drinking is a time management philosophy that balances patience and happiness).

Table . Characteristics comparison of alcohol videos with different popularities.

Dimensions, variable, and category	Values, n (%)	Low likes, n (%)	High likes, n (%)	Chi-square (*df*)	*P* value
Basic Information					
Durations				12.1 (2)	.002
0-30	340 (51.6)	184 (54.3)	156 (48.8)		
31-60	150 (22.8)	87 (25.7)	63 (19.7)		
>60	169 (25.6)	68 (20.1)	101 (31.6)		
Brand category				9.6 (1)	.002
New	299 (45.4)	134 (39.5)	165 (51.6)		
Traditional	360 (54.6)	205 (60.5)	155 (48.4)		
Alcohol category				161.6 (5)	<.001
Chinese liquor	118 (17.9)	42 (12.4)	76 (23.8)		
Beer	111 (16.8)	8 (2.4)	103 (32.2)		
Wine	103 (15.6)	80 (23.6)	23 (7.2)		
Imported liquor	74 (11.2)	53 (15.6)	21 (6.6)		
Fruit wine or preconditioning of cocktails	134 (20.3)	69 (20.4)	65 (20.3)		
Yellow or rice wine	119 (18.1)	87 (25.7)	32 (10.0)		
Content presentation					
Form of presentation				7.6 (4)	.11
Advertisement	281 (42.6)	147 (43.4)	134 (41.9)		
Short skit	255 (38.7)	122 (36.0)	133 (41.6)		
Film and television show excerpts	20 (3.0)	7 (2.1)	13 (4.1)		
Vlog	70 (10.6)	42 (12.4)	28 (8.8)		
Other	33 (5.0)	21 (6.2)	12 (3.8)		
Characters' drinking action				28.1 (3)	<.001
No characters and no drinking behavior	106 (16.1)	79 (23.3)	27 (8.4)		
Characters appear without drinking behaviors	184 (27.9)	92 (27.1)	92 (28.8)		
Characters appear with drinking-related behaviors	265 (40.2)	119 (35.1)	146 (45.6)		
Characters appear with drinking behaviors	104 (15.8)	49 (14.5)	55 (17.2)		
Scene setting					
Drinking alone				3.1 (1)	.08
Yes	99 (15.0)	59 (17.4)	40 (12.5)		
No	560 (85.0)	280 (82.6)	280 (87.5)		
Party				16.0 (1)	<.001
Yes	170 (25.8)	65 (19.2)	105 (32.8)		
No	489 (74.2)	274 (80.8)	215 (67.2)		

Dimensions, variable, and category	Values, n (%)	Low likes, n (%)	High likes, n (%)	Chi-square (*df*)	*P* value
Natural scenery				0.6 (1)	.45
Yes	112 (17.0)	54 (15.9)	58 (18.1)		
No	547 (83.0)	285 (84.1)	262 (81.9)		
Cultural or sports activities				3.5 (1)	.06
Yes	75 (11.4)	31 (9.1)	44 (13.8)		
No	584 (88.6)	308 (90.9)	276 (86.3)		
Brand and product appeal					
Brand elements				45.9 (1)	<.001
Yes	510 (77.4)	226 (66.7)	284 (88.8)		
No	149 (22.6)	113 (33.3)	36 (11.3)		
Product elements				2.0 (1)	.16
Yes	399 (60.5)	214 (63.1)	185 (57.8)		
No	260 (39.5)	125 (36.9)	135 (42.2)		
Intrinsic product features				48.2 (1)	<.001
Yes	254 (38.5)	174 (51.3)	80 (25.0)		
No	405 (61.5)	165 (48.7)	240 (75.0)		
Extended product features				28.0 (1)	<.001
Yes	161 (24.4)	112 (33.0)	49 (15.3)		
No	498 (75.6)	227 (67.0)	271 (84.7)		
Promotion strategy					
Product promotion				22.7 (3)	<.001
Key opinion leaders	53 (8.0)	14 (4.1)	39 (12.2)		
Cross-border brand cooperation	32 (4.9)	23 (6.8)	9 (2.8)		
Interaction with audience	12 (1.8)	3 (0.9)	9 (2.8)		
Not mentioned	562 (85.3)	299 (88.2)	263 (82.2)		
Cues refer to women’s interests				4.4 (1)	.04
Yes	241 (36.6)	111 (32.7)	130 (40.6)		
No	418 (63.4)	228 (67.3)	190 (59.4)		
Cues refer to teenagers’ interests				20.5 (1)	<.001
Yes	172 (26.1)	63 (18.6)	109 (34.1)		
No	487 (73.9)	276 (81.4)	211 (65.9)		
Emotion					
Emotional tone				3.0 (2)	.22
Positive	263 (39.9)	128 (37.8)	135 (42.2)		
Neutral	359 (54.5)	195 (57.5)	164 (51.3)		
Negative	37 (5.6)	16 (4.7)	21 (6.6)		
Culture					
Cultural appeal				40.4 (6)	<.001
Historical inheritance	65 (9.9)	43 (12.7)	22 (48.8)		

Dimensions, variable, and category	Values, n (%)	Low likes, n (%)	High likes, n (%)	Chi-square (*df*)	*P* value
Festival celebration	49 (7.4)	35 (10.3)	14 (6.9)		
Balance in life	45 (6.8)	8 (2.4)	37 (4.4)		
Ambition and striving	33 (5.0)	17 (5.0)	16 (19.7)		
Enjoy one's life	153 (23.2)	90 (26.5)	63 (11.6)		
Gift-giving culture	21 (3.2)	9 (2.7)	12 (5.0)		
Not mentioned	292 (44.5)	137 (40.4)	156 (3.8)		
Warning					
Age restriction				19.4 (1)	<.001
Yes	482 (73.1)	273 (80.5)	209 (65.3)		
No	177 (26.9)	66 (19.5)	111 (34.7)		
Health warnings				17.0 (1)	<.001
Drinking is harmful to health	8 (1.2)	6 (1.8)	2 (0.6)		
Please drink responsibly	68 (10.3)	50 (14.7)	18 (5.6)		
Not mentioned	583 (88.5)	283 (83.5)	300 (93.8)		

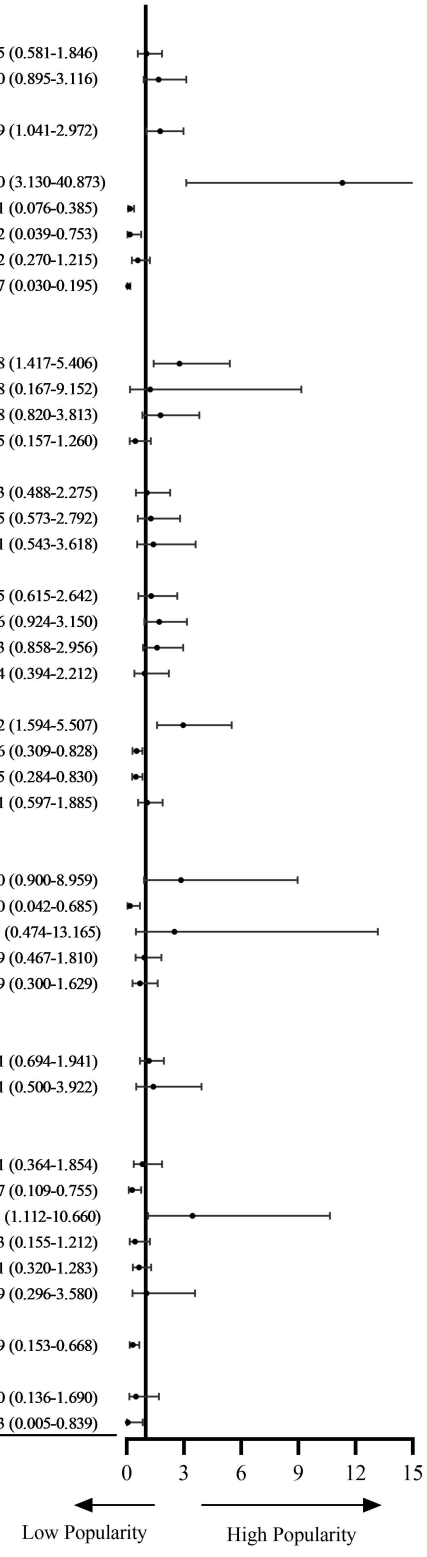
The Factors Associated With the Attractiveness of Alcohol Advertisements

Compared with videos from traditional brands, videos from new brands were more likely to receive more likes (OR 1.759, 95% CI 1.041 - 2.972). Compared with Chinese liquor videos, beer videos garnered more likes (OR 11.310, 95% CI 3.130 - 40.873), whereas wine, imported liquor, and yellow or rice wine videos received fewer likes (OR 0.171, 95% CI 0.076 - 0.385; OR 0.172, 95% CI 0.039 - 0.753; and OR 0.077, 95% CI 0.030 - 0.195).

Compared with direct advertisements, videos presented as short skits were more likely to receive likes (OR 2.768, 95% CI 1.417 - 5.406). Videos conveying the culture of "balance in life" received more likes (OR 3.442, 95% CI 1.112 - 10.660). Additionally, videos that displayed brand-related information were more attractive (OR 2.962, 95% CI: 1.594 - 5.507), whereas those showcasing product elements or intrinsic features garnered fewer likes (OR 0.506, 95% CI 0.309 - 0.828; and OR 0.485, 95% CI 0.284 - 0.830). Furthermore, videos with age restrictions and the health warning "Drinking is harmful to health" tended to receive fewer likes (OR 0.319, 95% CI 0.153 - 0.668; OR 0.063, 95% CI 0.005 - 0.839). The total Nagelkerke R^2 of this logistic regression was 0.561 (Figure 2).

Figure 2. Multivariate logistic regression of popularity of alcohol-related videos. OR: odds ratio.

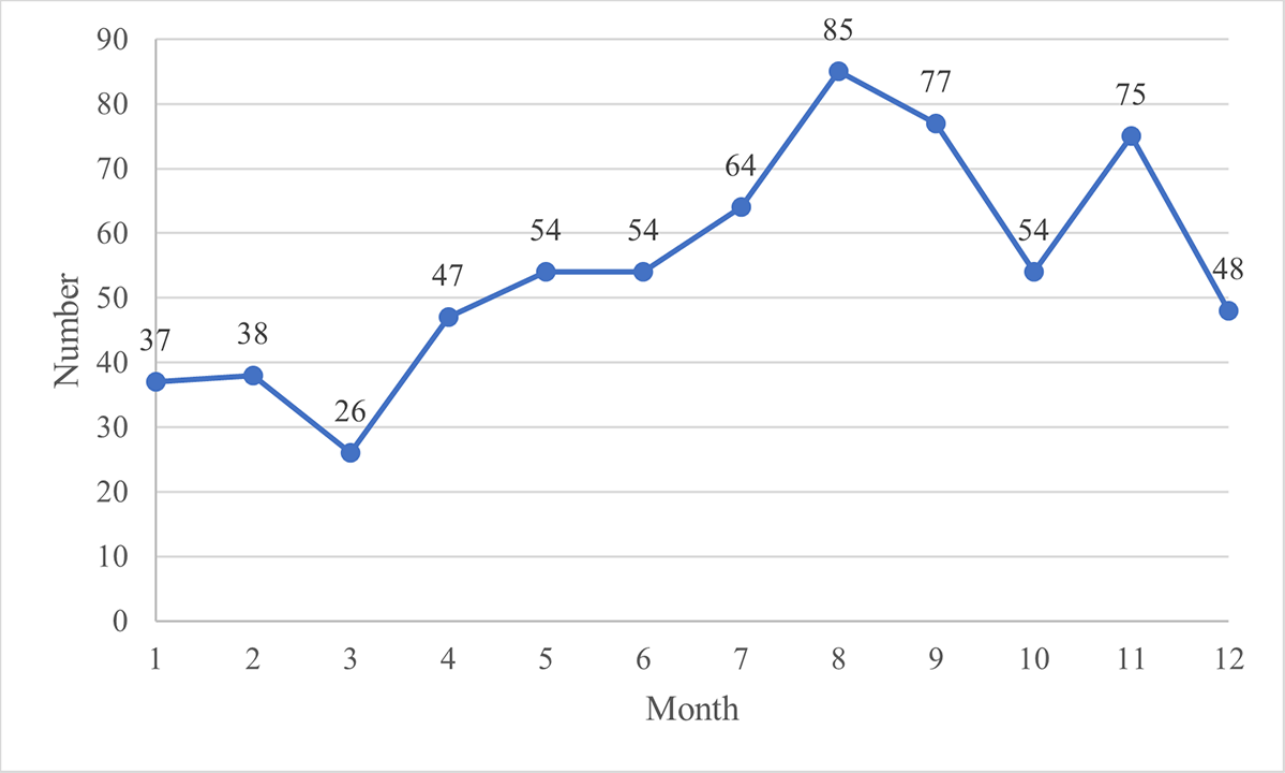
Dimension	Variable	Category	P	OR (95%CI)
Basic information				
	Duration	0s to 30s (Ref.)		
		31 to 60s	.91	1.035 (0.581-1.846)
		More than 60s	.11	1.670 (0.895-3.116)
	Brand category	Traditional (Ref.)		
		New	.04	1.759 (1.041-2.972)
	Alcohol category	Chinese liquor(Ref.)		
		Beer	<.001	11.310 (3.130-40.873)
		Wine	<.001	0.171 (0.076-0.385)
		Imported liquor	.02	0.172 (0.039-0.753)
		Fruit wine/ cocktails	.12	0.572 (0.270-1.215)
		Yellow/rice wine	<.001	0.077 (0.030-0.195)
Content presentation				
	Form of presentation	Advertisement (Ref.)		
		Short skit	.003	2.768 (1.417-5.406)
		Film and TV show excerpts	.84	1.238 (0.167-9.152)
		Vlog	.15	1.768 (0.820-3.813)
		Other	.13	0.445 (0.157-1.260)
	Characters' drinking action	No characters and no drinking behavior (Ref.)		
		Characters appear without drinking behaviors	.90	1.053 (0.488-2.275)
		Characters appear with drinking related behaviors	.56	1.265 (0.573-2.792)
		Characters appear with drinking behaviors directly	.49	1.401 (0.543-3.618)
Scene setting				
	Drinking alone	Yes vs No	.51	1.275 (0.615-2.642)
	Party	Yes vs No	.09	1.706 (0.924-3.150)
	Natural scenery	Yes vs No	.14	1.593 (0.858-2.956)
	Cultural and sports activities	Yes vs No	.88	0.934 (0.394-2.212)
Brand and product appeal				
	Brand elements	Yes vs No	.001	2.962 (1.594-5.507)
	Product elements	Yes vs No	.007	0.506 (0.309-0.828)
	Intrinsic product features	Yes vs No	.008	0.485 (0.284-0.830)
	Extended product features	Yes vs No	.84	1.061 (0.597-1.885)
Promotion strategy				
	Product promotion strategy	Not mentioned (Ref.)		
		Key opinion leaders	.08	2.840 (0.900-8.959)
		Cross-border brand cooperation	.01	0.170 (0.042-0.685)
		Interaction with audience	.28	2.497 (0.474-13.165)
	Elements attracting women	Yes vs No	.81	0.919 (0.467-1.810)
	Elements attracting teenagers	Yes vs No	.41	0.699 (0.300-1.629)
Emotion				
	Emotional tone	Neutral (Ref.)		
		Positive	.60	1.161 (0.694-1.941)
		Negative	.52	1.401 (0.500-3.922)
Culture				
	Cultural appeal	Not mentioned (Ref.)		
		Historical inheritance	.64	0.821 (0.364-1.854)
		Festival celebration	.01	0.287 (0.109-0.755)
		Balance in life	.03	3.442 (1.112-10.660)
		Ambition and striving	.11	0.433 (0.155-1.212)
		Enjoy one's life	.21	0.641 (0.320-1.283)
		Gift-giving culture	.96	1.029 (0.296-3.580)
Warning				
	Age restriction	Yes vs No	.002	0.319 (0.153-0.668)
	Health warnings	Not mentioned (Ref.)		
		Please drink responsibly	.25	0.480 (0.136-1.690)
		Drinking is harmful to health	.04	0.063 (0.005-0.839)



Temporal Trend of Alcohol Video Releases on Douyin

The period from July to September of that year was the peak season for alcohol advertisements on Douyin, with the highest monthly number reaching 85. In addition, a release peak for Douyin alcohol videos was recorded in November, with the number reaching 75 (Figure 3).

Figure 3. Distribution of Douyin alcohol advertisement release times.



Discussion

Principal Findings

To the best of our knowledge, this is the first study to analyze the thematic content of alcohol advertising on the Douyin platform. Through the analysis of 659 Douyin videos, we identified several alcohol marketing strategies, including diverse forms of presentation, highlighting the brand elements and incorporating cultural elements into the videos. In addition, we explored the factors influencing the popularity of those videos and reported that the use of several marketing approaches was positively correlated with the attractiveness of the videos. All these alcohol marketing practices may accelerate public alcohol consumption through enhancing alcogenic environment. These findings emphasize the urgent need for strong policy formulation and enforcement to reduce the negative influence of alcohol marketing on social media platforms.

Alcohol Types and Temporal Trends

According to the results, beer advertisements were more popular than Chinese liquor advertisements, whereas wine, imported liquor, and rice or yellow wine advertisements received less attention. We speculate that this is due to the high overlap between the characteristics of the beer-consuming audience and Douyin users. According to the Douyin user portrait report, the main audience of Douyin consists of young people aged 19 to 30 years from non–first-tier cities [28]. Chinese youth primarily prefer beer, followed by Chinese liquor, whereas wine, imported liquor, and yellow wine are far less popular [29]. The temporal trend of Douyin alcohol videos also reflects young Chinese people's preferences. Unlike the traditional peak season of Chinese liquor during the Spring Festival season (December to February) [30], alcohol advertisements on Douyin are heavily launched during the summer vacation period (July to September) as well as the month of the Double 11 Shopping Festival to cater to young people's demand for beer consumption. As in previous studies, all these factors indicate that social media platforms such as Douyin have become popular alcohol marketing channels targeting youth [31-33].

Alcohol Marketing Strategies on Douyin

The first alcohol marketing strategy on the Douyin platform is experiential marketing, a prime example of which is the utilization of short skits [34]. Unlike traditional advertising, these skits occur within an entertainment context, becoming part of an experience that immerses the viewer in the storyline [35]. This approach may blur the lines between audiences' personal lives and commercial messages while frequently depicting alcohol-related behaviors, which may encourage higher drinking frequency among viewers [31,36,37]. Another marketing strategy is brand marketing, which emphasizes brand elements such as the logo or mascot, rather than simple introductions of products. These elements could convey brand emotions and values, prompting people to hold a positive attitude toward a brand or product, thus increasing the probability of purchasing [38,39]. Brand marketing is not only common in China but has also been validated by previous studies as one of the prevalent strategies in global alcohol marketing [40].

The third strategy is collaborative marketing, in which the involvement of celebrities becomes the main strategy. Positive characteristics associated with the celebrity can be transferred to the product. In particular, the celebrity endorsement of young people could increase their recollection of drinking images [32,41]. In addition, many alcohol brands choose to collaborate with other brands or activities, including the National Basketball

Association and music festivals, which could also broaden their audience beyond traditional alcohol consumers. The last strategy is cultural marketing. Many alcohol brands incorporate cultural elements into their promotions [42]. Similar to previous studies, this is also the most commonly used marketing strategy in alcohol advertisements on Chinese television [43]. Previous studies have demonstrated that adapted cultural value appeals are more persuasive and attractive in advertisements [44]. On the other hand, alcohol advertising could further strengthen the "alcohol culture," enhance the alcogenic environment, and ultimately promote alcohol consumption.

This study extends the application of the AIDA model to alcohol advertising on the Douyin platform. Multivariate analysis revealed that when audiences are exposed to alcohol videos using experiential, brand, collaborative, and cultural marketing strategies, these videos are more likely to gain audience likes, thereby facilitating the transition from "attention" to "interest." To prevent the transition of audience attention to purchasing action, reducing the use of these marketing strategies in alcohol videos is crucial, thereby limiting the formation of "interest."

Age Restriction and Health Warning

Notably, many videos include cues related to teenagers' interests, such as idols, cartoons, and e-sports. Although there is currently no regulation explicitly requiring alcohol advertisements to include age restriction warnings, Douyin's platform policy clearly mandates that alcohol advertisements must contain warnings such as "Users under the age of 18 are prohibited from purchasing this product." [45] However, not all videos contain those age restriction signs. Although the presence of age restrictions may reduce the number of likes, the actual effectiveness of these restrictions remains unclear. It is generally easy to access social media platforms, and even children younger than 13 years can "legally" use social media and view alcohol advertisements, despite these age restrictions in place [46]. It is urgent to standardize the "teenage mode" on these social media platforms to mitigate the influence of alcohol marketing on teenagers.

Moreover, the lack of health warnings on social media platforms is a serious issue. Only 11.5% of the videos include any form of health warning, and most of these are ambiguous messages, "drink responsibly." In fact, "drink responsibly" messages are associated with increased alcohol consumption [47]. This is because such messages tend to enhance a prodrinking social norm. When "responsible drinking" messages are placed in alcohol advertisements, the alcohol industry can give the impression of fulfilling corporate responsibility without decreasing sales [48,49]. As reported in previous studies, only explicit health warnings that inform consumers about the carcinogenic effects of alcohol have a significant effect, rather than ambiguous messages [49]. Consequently, the 2022‐2030 WHO Global Alcohol Action Plan has called for ensuring that the labeling of alcoholic beverages is appropriate [50] and that essential health protection information is displayed.

Policy Implications

Although China's Advertising Law stipulates that all alcohol advertisements must not encourage drinking, depict drinking behaviors, or emphasize the functional benefits of alcohol consumption, and additionally prohibits the dissemination of alcohol advertisements through mass media targeting minors [51], alcohol marketing remains highly prevalent on social media platforms such as Douyin, where minors can still be easily exposed to such content. This suggests that deficiencies in both the legal provisions on alcohol advertising and their enforcement remain.

With respect to laws governing alcohol advertising, first, the definition of alcohol advertising should be refined to cover not only direct advertisements but also embedded stealth advertising, such as short skits and placements. Thus, legislative restrictions should include a comprehensive ban on all forms of alcohol commercial communication, recommendation, or activity. While these factors may not directly encourage drinking, they could help enhance the alcogenic environment. Moreover, the definition of "media targeting minors" remains ambiguous, making it possible for minors to be exposed to alcohol advertisements through general-audience channels, especially social media [43]. It is imperative to either explicitly prohibit alcohol advertising on social media or establish a robust "teenage mode." With respect to age restriction warnings and health warnings for alcohol advertising, mandatory regulations should not be limited to platforms such as Douyin but must be incorporated into advertising laws and strictly enforced. Finally, health warnings should be provided to avoid vague expressions such as "drink responsibly" and instead provide specific examples of the health risks associated with alcohol consumption.

With respect to law enforcement, the responsibility for monitoring and enforcing the existing regulations on advertising in China lies primarily with only the current commercial administrative department. In the future, other administrative departments, including health, food, and drug departments, could also participate to establish a comprehensive enforcement network [21]. Previous studies have demonstrated that the low penalties in the case of violation and the lack of effective detection are also key factors hindering its effective enforcement [21]. Enhancing penalty severity, improving regulatory channels, and encouraging public participation in supervision can significantly strengthen the enforcement of the law. With comprehensive legislation and implementation of restrictions, reducing the alcogenic environment caused by alcohol marketing is among the most cost-effective ways.

Limitations

The data collection period for this study (2021) may present certain limitations. With the rapid evolution of the digital marketing environment, the promotional strategies of alcohol brands on social media platforms, such as Douyin, may have become more covert and innovative. In recent years, emerging approaches such as algorithm-driven personalized content placement, interactions with virtual spokespersons, and cross-platform integrated marketing campaigns have increased in prevalence. These strategies are often integrated more deeply into users' everyday browsing experiences, further blurring the boundary between commercial promotion and organic content. Future monitoring of alcohol advertising should focus on such

covert marketing techniques to mitigate the potential impact of alcogenic environments.

There are also several limitations of this study. First, this study only gathered alcohol advertisements from Douyin over a 1-year period, failing to gather data over an extended period to assess the changing trends in marketing strategies over the years. Future studies could conduct longer-term longitudinal research to analyze temporal trends of marketing strategies in alcohol advertising. Second, like counts were used as the only variable representing popularity. The specific expressions and interactions of the audience in the comment section were not included and need further exploration. Despite these limitations, this study lays a foundation for future research on alcohol marketing content on Chinese social media platforms and provides evidence for strengthening the regulation of alcohol marketing on social media platforms.

Conclusions

In conclusion, this study summarized several alcohol marketing strategies on the Douyin platform, including experiential marketing, brand building, cross-border collaboration, and cultural connection. These strategies may enhance video attractiveness and appeal to teenagers, serving as key factors in transforming “attention” into “interest” (2 basic elements in the AIDA model) of alcohol advertisements. However, effective age restrictions and explicit health warnings are rarely shown in these alcohol-related Douyin videos. Urgent actions should include closing existing legal loopholes, such as refining the definition of alcohol advertising, strengthening protections for minors, and requiring specific health warnings, along with enhancing multiagency collaboration and imposing stricter penalties to decrease the alcogenic environment.

Acknowledgments

The authors are grateful to all study participants for their participation.

Funding

This work was supported by the National Natural Science Foundation of China (72474052).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

PZ, YZ, and LZ conceptualized and designed the study. YZ and LZ were responsible for data curation and formal analysis under the supervision of PZ. YZ wrote the original draft. CQ, WG, WZ, and PZ contributed to reviewing and editing the manuscript. All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The primary coding book of alcohol videos on Douyin platform.
[DOCX File, 19 KB - infodemiology_v6i1e74221_app1.docx]

References

1. Alcohol. WHO. URL: https://www.who.int/news-room/fact-sheets/detail/alcohol [accessed 2025-09-09]
2. A health promotion approach for reducing youth exposure to alcogenic environments: snapshot series on alcohol policies and practice. WHO. 2023. URL: https://www.who.int/publications/i/item/9789240073296 [accessed 2025-09-09]
3. Hill KM, Foxcroft DR, Pilling M. “Everything is telling you to drink”: understanding the functional significance of alcogenic environments for young adult drinkers. Addict Res Theory 2018 Nov 2;26(6):457-464. [doi: 10.1080/16066359.2017.1395022]
4. Huckle T, Huakau J, Sweetsur P, Huisman O, Casswell S. Density of alcohol outlets and teenage drinking: living in an alcogenic environment is associated with higher consumption in a metropolitan setting. Addiction 2008 Oct;103(10):1614-1621. [doi: 10.1111/j.1360-0443.2008.02318.x] [Medline: 18821871]
5. Murphy A, Roberts B, Ploubidis GB, Stickley A, McKee M. Using multi-level data to estimate the effect of an “alcogenic” environment on hazardous alcohol consumption in the former Soviet Union. Health Place 2014 May;27:205-211. [doi: 10.1016/j.healthplace.2014.02.015] [Medline: 24662529]
6. Petticrew M, Shemilt I, Lorenc T, et al. Alcohol advertising and public health: systems perspectives versus narrow perspectives. J Epidemiol Community Health 2017 Mar;71(3):308-312. [doi: 10.1136/jech-2016-207644] [Medline: 27789756]

7. Stockings E, Bartlem K, Hall A, et al. Whole-of-community interventions to reduce population-level harms arising from alcohol and other drug use: a systematic review and meta-analysis. Addiction 2018 Nov;113(11):1984-2018. [doi: 10.1111/add.14277] [Medline: 29806876]
8. Moreno MA, Christakis DA, Egan KG, Brockman LN, Becker T. Associations between displayed alcohol references on Facebook and problem drinking among college students. Arch Pediatr Adolesc Med 2012 Feb;166(2):157-163. [doi: 10.1001/archpediatrics.2011.180] [Medline: 21969360]
9. Monitoring and restricting digital marketing of unhealthy products to children and adolescents. WHO. URL: https://www.who.int/europe/activities/monitoring-and-restricting-digital-marketing-of-unhealthy-products-to-children-and-adolescents [accessed 2025-09-09]
10. Lobstein T, Landon J, Thornton N, Jernigan D. The commercial use of digital media to market alcohol products: a narrative review. Addiction 2017 Jan;112(S1):21-27. [doi: 10.1111/add.13493] [Medline: 27327239]
11. Bouckley B. Anheuser-busch: facebook beats any US broadcast network for consumer reach. Beverage Daily. 2013. URL: https://www.beveragedaily.com/Article/2013/11/12/Anheuser-Busch-Facebook-consumer-reach-beats-any-US-broadcast-network/ [accessed 2025-09-09]
12. Cranwell J, Opazo-Breton M, Britton J. Adult and adolescent exposure to tobacco and alcohol content in contemporary YouTube music videos in Great Britain: a population estimate. J Epidemiol Community Health 2016 May;70(5):488-492. [doi: 10.1136/jech-2015-206402] [Medline: 26767404]
13. Manthey J, Shield KD, Rylett M, Hasan OSM, Probst C, Rehm J. Global alcohol exposure between 1990 and 2017 and forecasts until 2030: a modelling study. Lancet 2019 Jun 22;393(10190):2493-2502. [doi: 10.1016/S0140-6736(18)32744-2] [Medline: 31076174]
14. Peng C. Digital transformation and new retail model exploration of liquor enterprise marketing [Article in Chinese]. Enterp Reform Manag 2021(17):112-113. [doi: 10.13768/j.cnki.cn11-3793/f.2021.1744]
15. China liquor consumption insight report 2021. Tencent News. 2021. URL: https://jiu.ifeng.com/c/89c8yf4DYxr [accessed 2025-09-09]
16. 2024 Douyin user portrait and audience analysis. TikTok. 2025. URL: https://www.jingliren.org/rdxx/hlw/311578.html [accessed 2025-09-09]
17. #Chinese liquor. Tiktok. URL: https://www.douyin.com/hashtag/1583910072238157 [accessed 2025-09-09]
18. White book on Chinese liquor consumption 2023. Chinese Alcoholic Drinks Association. 2023. URL: https://file.tencentads.com/web/pdf/index/f90b46878551cdba [accessed 2025-09-09]
19. Tristanto TA, Hurriyati R. AIDA model as a marketing strategy to influence consumer buying interest in the digital age. BIRCI 2023;4:12575-12586. [doi: 10.33258/birci.v4i4.3319]
20. Chan RHW, Dong D, Chong MKC, et al. Alcohol social media marketing and drinking behaviors among Chinese young adults: Mediation by drinking expectancies. Drug Alcohol Depend 2025 Oct 1;275:112818. [doi: 10.1016/j.drugalcdep.2025.112818] [Medline: 40769088]
21. Ji N, Xu Q, Zeng X, Casswell S, Bai Y, Liu S. Alcohol advertising exposure and drinking habits among Chinese adolescents in 2021: a national survey. Am J Public Health 2024 Aug;114(8):814-823. [doi: 10.2105/AJPH.2024.307680] [Medline: 38870435]
22. China liquor industry development research report 2021. iiMedia research. 2021. URL: https://baijiahao.baidu.com/s?id=1718016690969263242&wfr=spider&for=pc [accessed 2025-09-09]
23. JD Global. URL: https://global.jd.com/ [accessed 2025-12-31]
24. The rise of Douyin and TikTok: the commercially accelerating short video empire after 2020. Vzkoo. 2024. URL: https://www.vzkoo.com/read/2024111213885f4bdbc1f915258ac52d.html [accessed 2025-12-05]
25. Barker AB, Bal J, Murray RL. A content analysis and population exposure estimate of Guinness branded alcohol marketing during the 2019 Guinness Six Nations. Alcohol Alcohol 2021 Aug 30;56(5):617-620. [doi: 10.1093/alcalc/agab039] [Medline: 34080614]
26. Atkinson AM, Sumnall H, Meadows B. “We’re in this together”: a content analysis of marketing by alcohol brands on Facebook and Instagram during the first UK Lockdown, 2020. Int J Drug Policy 2021 Dec;98:103376. [doi: 10.1016/j.drugpo.2021.103376] [Medline: 34364199]
27. Notice on issuing the measures for ethical review of human life science and medical research. China’s National Health Commission, Ministry of Education, and Ministry of Science and Technology. 2023. URL: https://www.gov.cn/zhengce/zhengceku/2023-02/28/content_5743658.htm [accessed 2025-12-22]
28. TikTok user profile report 2024. CSDN. 2024. URL: https://blog.csdn.net/2301_76223496/article/details/137100176?spm=1001.2101.3001.6650.3 [accessed 2025-09-09]
29. GE Lianying LR, Ning LI, Jing ZHU, Dan XU. Investigation and analysis on drinking culture of youth [Article in Chinese]. Modern Food 2022;28(21):224-228. [doi: 10.16736/j.cnki.cn41-1434/ts.2022.21.053]
30. Mid-2025 research report on China’s baijiu market. : Chinese Alcoholic Drinks Association; 2025 URL: https://assets.kpmg.com/content/dam/kpmg/cn/pdf/zh/2025/06/mid-term-research-report-on-the-chinese-baijiu-market-2025.pdf [accessed 2025-09-09]

31. Barry AE, Padon AA, Whiteman SD, et al. Alcohol advertising on social media: examining the content of popular alcohol brands on Instagram. Subst Use Misuse 2018 Dec 6;53(14):2413-2420. [doi: 10.1080/10826084.2018.1482345] [Medline: 29889647]
32. Hendriks H, Wilmsen D, van Dalen W, Gebhardt WA. Picture me drinking: alcohol-related posts by Instagram influencers popular among adolescents and young adults. Front Psychol 2019;10:2991. [doi: 10.3389/fpsyg.2019.02991] [Medline: 32038379]
33. Guégan E, Zenone M, Mialon M, Gallopel-Morvan K. #Bartender: portrayals of popular alcohol influencer's videos on TikTok©. BMC Public Health 2024 May 23;24(1):1384. [doi: 10.1186/s12889-024-18571-1] [Medline: 38783213]
34. Mart SM. Alcohol marketing in the 21st century: new methods, old problems. Subst Use Misuse 2011;46(7):889-892. [doi: 10.3109/10826084.2011.570622] [Medline: 21599504]
35. Cowley E, Barron C. When product placement goes wrong: the effects of program liking and placement prominence. J Advert 2008 Apr;37(1):89-98. [doi: 10.2753/JOA0091-3367370107]
36. Eagle L, Dahl S. Product placement in old and new media: examining the evidence for concern. J Bus Ethics 2018 Feb;147(3):605-618. [doi: 10.1007/s10551-015-2955-z]
37. D'Angelo J, Kerr B, Moreno MA. Facebook displays as predictors of binge drinking: from the virtual to the visceral. Bull Sci Technol Soc 2014;34(5-6):159-169. [doi: 10.1177/0270467615584044] [Medline: 26412923]
38. Beukeboom CJ, Kerkhof P, De Vries M. Does a virtual like cause actual liking? How following a brand's Facebook updates enhances brand evaluations and purchase intention. J Interact Mark 2015 Nov;32(1):26-36. [doi: 10.1016/j.intmar.2015.09.003]
39. Toldos-Romero MP, Orozco-Gómez MM. Brand personality and purchase intention. Eur Bus Rev 2015 Aug 10;27(5):462-476. [doi: 10.1108/EBR-03-2013-0046]
40. Jernigan D, Ross CS. The alcohol marketing landscape: alcohol industry size, structure, strategies, and public health responses. J Stud Alcohol Drugs Suppl 2020 Mar(19):13-25. [doi: 10.15288/jsads.2020.s19.13] [Medline: 32079559]
41. Jernigan DH, Padon A, Ross C, Borzekowski D. Self-reported youth and adult exposure to alcohol marketing in traditional and digital media: results of a pilot survey. Alcohol Clin Exp Res 2017 Mar;41(3):618-625. [doi: 10.1111/acer.13331] [Medline: 28219114]
42. Brodmerkel S, Carah N. Alcohol brands on Facebook: the challenges of regulating brands on social media. J Public Aff 2013 Aug;13(3):272-281. [doi: 10.1002/pa.1466]
43. Tang Y, Lei N, Hu D, et al. Estimated exposure to televised alcohol advertisements among children and adolescents. JAMA Netw Open 2025 Jul 1;8(7):e2521819. [doi: 10.1001/jamanetworkopen.2025.21819] [Medline: 40674047]
44. Hornikx J, Janssen A, O'Keefe DJ. Cultural value adaptation in advertising is effective, but not dependable: a meta-analysis of 25 years of experimental research. Int J Bus Commun 2023. [doi: 10.1177/23294884231199088]
45. 【Alcoholic beverages】product release and promotion guidelines. The Douyin E-commerce Operations Team. 2025. URL: https://school.jinritemai.com/doudian/web/article/aHSh8Zmd66rx#:~:text [accessed 2025-12-05]
46. Barry AE, Johnson E, Rabre A, Darville G, Donovan KM, Efunbumi O. Underage access to online alcohol marketing content: a YouTube case study. Alcohol Alcohol 2015 Jan;50(1):89-94. [doi: 10.1093/alcalc/agu078] [Medline: 25411395]
47. Moss AC, Albery IP, Dyer KR, et al. The effects of responsible drinking messages on attentional allocation and drinking behaviour. Addict Behav 2015 May;44:94-101. [doi: 10.1016/j.addbeh.2014.11.035] [Medline: 25577316]
48. Maani Hessari N, Petticrew M. What does the alcohol industry mean by "Responsible drinking"? A comparative analysis. J Public Health 2018 Mar 1;40(1):90-97. [doi: 10.1093/pubmed/fdx040] [Medline: 28398571]
49. Davies E, Lewin J, Field M. Am I a responsible drinker? The impact of message frame and drinker prototypes on perceptions of alcohol product information labels. Psychol Health 2024 Aug;39(8):1005-1022. [doi: 10.1080/08870446.2022.2129055] [Medline: 36190181]
50. Global alcohol action plan 2022-2030. WHO. 2024. URL: https://www.who.int/publications/i/item/9789240090101 [accessed 2025-09-09]
51. Advertisement Law of the People's Republic of China. State Administration for Market Regulation. 2021. URL: https://www.samr.gov.cn/zw/zfxxgk/fdzdgknr/fgs/art/2023/art_5474cf75173c45d6a0379730fb4e8d97.html [accessed 2025-09-09]

Abbreviations

AIDA: attention-interest-desire-action
OR: odds ratio
WHO: World Health Organization

Edited by T Mackey; submitted 20.Mar.2025; peer-reviewed by A Pal, Y Yang; accepted 09.Dec.2025; published 06.Jan.2026.

Please cite as:
Zhao Y, Zhang L, Qian C, Guo W, Zhu W, Zheng P
Marketing Strategies and Factors Influencing the Popularity of Alcohol Videos from Official Brand Accounts on Douyin: Content Analysis Study
JMIR Infodemiology 2026;6:e74221
URL: https://infodemiology.jmir.org/2026/1/e74221
doi:10.2196/74221

Original Paper

Using Artificial Intelligence Methods to Evaluate the Effect of the National Cytomegalovirus Awareness Month on the Content and Sentiment of Social Media Posts: Infodemiology Study

Tracy R Rosebrock[1], PhD, MPH; Zhen Yang[2], MS, MScAc; Lauren D'Arco[1], BA; Tapan Pathak[2], MS; Rebecca Vislay-Wade[2], MA, MS, PhD; Karen Fowler[3], MS, PhD; John Diaz-Decaro[2], MS, PhD; Colin Kunzweiler[2], MS, PhD

[1]Department of Health Science, Department of Biology, School of Arts and Sciences, Stonehill College, North Easton, MA, United States

[2]Moderna Therapeutics (United States), Cambridge, MA, United States

[3]Department of Pediatrics, Heersink School of Medicine, University of Alabama at Birmingham, Birmingham, AL, United States

Corresponding Author:
Tracy R Rosebrock, PhD, MPH
Department of Health Science, Department of Biology
School of Arts and Sciences
Stonehill College
Shields Science Center
320 Washington Street
North Easton, MA, 02357
United States
Phone: 1 5085651097
Email: trosebrock@stonehill.edu

Abstract

Background: The month of June has been recognized as the National Cytomegalovirus (CMV) Awareness Month since 2011 in the United States. Established by government resolution, the goal is to increase awareness and reduce the incidence of congenital CMV infection, a leading cause of preventable birth defects and developmental disabilities. Social media is a powerful tool to support public health by making health information easily accessible. With an estimated 246 million users in the United States and more than half of adults seeking health information through such platforms, social media offers an unparalleled opportunity to promote CMV awareness and prevention.

Objective: This study aimed to evaluate social media messaging before, during, and after the National CMV Awareness Month to assess how the campaign influenced messaging patterns and sentiment related to specific CMV health topics.

Methods: Publicly available posts on Twitter/X from May to August 2023 that contained at least one of the five most used CMV-related hashtags were collected using a media monitoring platform. The dataset was preprocessed using a customized Bidirectional Encoder Representations from Transformers tokenizer and a language detection package to remove irrelevant and non-English posts. Validated and artificial intelligence (AI) methods (Cohen κ=0.69) were used to determine the thematic content of posts (N=14,900), such as awareness and prevention messaging, and to characterize the sentiment. Changes in post characteristics were measured in relation to the National CMV Awareness Month.

Results: CMV-relevant post volume increased by 55% during the campaign month and returned to precampaign levels in July. Overall, academic/university researchers were the most frequent authors, pediatrics was the most frequent population discussed, and vaccines were the most frequently mentioned prevention. Significant associations were observed between the month of post publication and the target audience (χ^2_2=144.3, *P*<.001), awareness or prevention messaging (χ^2_2=107.8, *P*<.001), and post sentiment (χ^2_4=163.6, *P*<.001). The intended audience of posts shifted toward the general population from scientists/health care professionals during the campaign month (adjusted Pearson residuals, *P*=.009). Awareness messaging increased in June 2023, particularly in relation to CMV transmission and disease burden, while prevention messaging decreased (adjusted Pearson residuals, *P*=.008). Finally, although posts were generally neutral in sentiment, a significant shift occurred toward a positive sentiment during the campaign month (adjusted Pearson residuals, *P*=.006), a sentiment that was more likely to engage the user (Kruskal-Wallis; χ^2_2=194.31, *P*<.001).

Conclusions: The National CMV Awareness Month in 2023 shifted the digital CMV conversation toward public-facing messaging and raised awareness efforts. Although posts related to CMV prevention generally conveyed a positive sentiment, prevention messaging declined during the campaign. These findings highlight opportunities for future CMV social media initiatives to balance awareness with prevention through evaluation and strategic design using AI models to strengthen CMV public health communication and engagement.

(JMIR Infodemiology 2026;6:e80922) doi:10.2196/80922

KEYWORDS

cytomegalovirus; social media; public health, health communication; sentiment analysis; artificial intelligence

Introduction

Background

Human cytomegalovirus (CMV) is a ubiquitous beta-herpesvirus with an estimated seroprevalence of 83% worldwide and 63% in the United States and a disease burden that disproportionally impacts disadvantaged and minoritized communities [1-4]. The virus is transmitted through direct contact with infectious body fluids, through organ/stem cell transplants, and transplacentally, which can result in congenial CMV infection [5,6]. Infection with CMV is lifelong and is typically subclinical in healthy individuals, though CMV may silently contribute to chronic conditions, such as cardiovascular disease [7], cognitive decline [8], neurologic disorders [9,10], anxiety and depression [11], immunosenescence [12], Guillain-Barré syndrome [13], certain malignancies (eg, glioblastoma multiforme) [14], and all-cause mortality [15]. In those who are immunocompromised, CMV infection can lead to acute disease and death [16,17].

When CMV is transmitted from mother to fetus during pregnancy, it is referred to as congenital cytomegalovirus (cCMV). In the United States, 1 in 200 infants is born with cCMV [18]. The majority of infants born with cCMV are asymptomatic without recognizable signs or symptoms at birth, while a subset (~10%) may be impacted by a wide range of signs and symptoms at birth [19]. Regardless of the presence or absence of symptoms at birth, all infants born with cCMV are at risk for developing long-term sequelae, such as sensorineural hearing loss, vision impairment, cerebral palsy, and developmental delays [20]. CMV is also a recognized cause of stillbirth and intrauterine fetal demise [21-24]. Despite the significant burden of cCMV [25,26], just 13% of women are aware of CMV, cCMV, or simple hygiene prevention practices that can reduce risk [27]. Given the limited effectiveness of pharmaceutical interventions and the lack of a licensed vaccine to prevent infection, public health education is of critical importance.

In 2011, the US Senate issued a resolution declaring June to be the National CMV Awareness Month, with explicit objectives to raise awareness of the dangers of CMV and to reduce the incidence of cCMV infections through education [28]. Today, eHealth, or the delivery of health information digitally, is recognized as an important frontier for public health education [29]. Social media is an important tool for public health education in the United States, where ~246 million people, or 73% of the population, are active users [30]. Among US adults, 55% report using social media at least occasionally to seek health information, and for the ~95 million US users of Twitter/X, the average daily use is greater than 30 minutes per day on the platform [31,32]. These platforms offer a valuable opportunity to disseminate accurate information and raise awareness about cCMV and prevention.

The National CMV Awareness Month is primarily promoted by the National Cytomegalovirus Foundation (NCMVF), advocacy groups of families affected by cCMV, and state and national public health agencies [33,34]. Shared goals are to increase public awareness and educate on prevention behaviors, while advocacy groups also promote expanded CMV screening and research. The awareness campaign is visible on social media through educational graphics, family stories, and hashtags, such as #stopCMV and #CMVawareness, to spread information and engage the public. Although CMV-specific campaign activity is described by advocacy organizations, prior public health infodemiological research has not systematically evaluated how these campaigns function on social media platforms.

Natural language processing (NLP) and sentiment analysis are novel, powerful, and essential tools in public health for monitoring public opinions toward health-related topics and identifying potential areas and concerns. For example, recent research has used Bidirectional Encoder Representations from Transformers (BERT) to analyze social media posts to understand public opinions toward the impact of the COVID-19 pandemic on social life [35] and sentiments expressed during an outbreak in Uganda [36]. Despite coordinated social media campaigns dedicated to CMV awareness, to date, sentiment analysis has not been conducted regarding this disease.

Beyond BERT-based approaches, recent infodemiology studies have increasingly incorporated large language models (LLMs) to analyze high-volume Twitter/X data for public health insight. For example, artificial intelligence (AI) models have been used to classify tweets related to conjunctivitis outbreaks and estimate epidemic signals, and LLM-driven sentiment and substance-use detection models have been applied to opioid-related social media data [37,38]. Such studies demonstrate the expanding role of generative and transformer-based models in characterizing eHealth discourse, providing methodological precedent for applying LLMs to evaluate public health messaging on Twitter/X.

Although survey-based studies have assessed CMV awareness and attitudes toward prenatal and neonate CMV screening [39,40], formal infodemiology or eHealth investigations, particularly those evaluating social media messaging related to CMV awareness and prevention, have not been conducted. This study represents the first systematic evaluation of CMV-related

discourse on social media, characterizing Twitter/X messaging during the National CMV Awareness Month and assessing how this communication aligns with goals set forth by the US Senate and public health stakeholders. Although prior infodemiology research has analyzed other public health campaigns or disease-related discourse using transformer models and LLMs, no studies have examined CMV or cCMV infections, and none have assessed a nationally recognized awareness-month campaign. By leveraging an LLM to classify, summarize, and evaluate campaign-related tweets, this study used emerging LLM-based methods to fill an important gap in the literature on digital CMV education and awareness.

Study Objective

The aim of this study was to evaluate the messaging of CMV-related posts on Twitter/X before, during, and after the National CMV Awareness Month in June 2023 (ie, May-August 2023).

Methods

Data Aggregation

Investigators collected social media posts from the Twitter/X platform (Twitter became X on July 23, 2023) for May-August 2023, representing the months immediately preceding and following the National CMV Awareness Month in June using Keyhole, a subscription-based, commercial social listening and analytics platform. Keyhole provides real-time tracking, historical backfill, and reach/impression estimates based on post volume and author follower counts.

To be eligible for this initial download, social media posts must have included one or more of the following five hashtags: #stopCMV, #cCMV, #CMV, #CMVawareness, and #cytomegalovirus. No additional filters were applied. These hashtags were identified by two independent reviewers, who manually assessed Twitter/X posts referencing "cytomegalovirus" or "CMV" using the site's search function. The most recent posts were assessed weekly from March 1 to April 30, 2023. From relevant posts that included hashtags, the five hashtags used in this study were the most frequently used and captured all posts in the assessment window that used hashtags.

In addition to the social media post text, user information (eg, username, biography, number of followers, and location) and dissemination-related metrics (eg, number of reposts, number of likes, and number of comments) were downloaded for each post. More than 30,000 social media posts were downloaded based on the aforementioned criteria.

Data Preparation

Non-English language posts were detected using the language-detecting Python package, *langdetect*, and removed during preliminary data-cleaning procedures. Among the more than 20,000 English-language social media posts that remained, posts were first cleaned to remove noise and other nuisance text (eg, URLs, emojis, and hashtag [#] and mention [@] symbols). A fine-tuned text classifier was next applied to identify relevant social media posts and exclude irrelevant social media posts from analyses [41]. Briefly, a customized BERT tokenizer was used to complete word-level tokenization of all social media posts. Reading social media posts from both left and right, the BERT tokenizer can understand the language context and flow of words based on a given word's surroundings. By default, BERT tokenizes, or breaks down, words into subword units. For example, the term "congenital" may be segmented into the subword units "con," "##gen," and "##ital" (the "##" prefix indicates that the subword is a continuation of the previous token unit). However, during exploratory data analysis, it was noticed that certain keywords, and in particular abbreviations, were segmented into individual characters (eg, "CMV" was segmented into "C," "##M," and "##V). To ensure meaningful pretraining, the BERT tokenizer was customized by manually adding a dictionary of domain-specific words (eg, "congenital," "cCMV," and "CMV") that would be identified as complete words, as opposed to subword units. We compared model performance using the default bert-base-cased tokenizer versus the customized tokenizer, each fine-tuned on the same manually labeled dataset. Both models achieved comparable accuracy (95%), indicating that the custom dictionary did not materially affect overall performance. Additional methodological details are provided in Multimedia Appendix 1.

Descriptive Analyses

Numerous descriptive analyses were conducted. The date of social media posts was analyzed to describe the number of posts per month for the May-August 2023 period. The location of each relevant, English-language social media post (when such information was available) was summarized at the country level (data not shown), as well as at the state level for posts originating in the United States. User categories (eg, university/academic researchers, news/journal/public health/education, and physicians/health care/hospitals) were summarized for the top 20 users with the most social media posts or most followers during the study period. The biography data supplied in the author's Twitter/X biosketch were used to determine their user category, or if a biography statement was not provided, the author was located using a web search and categorized using available data. Metrics of impact, including the number of followers, reposts, and "likes," were also summarized.

Theme (Aspect) Classification

Following the identification of relevant social media posts, investigators developed a master prompt for Moderna's internal ChatGPT AI tool (mChat), as shown in Multimedia Appendix 2, to annotate the specific aspects contained within each post. mChat serves as a pass-through to the OpenAI application programming interface (API), does not modify model behavior, and directly calls the OpenAI API using company credentials.

Aspects were first prespecified by two independent investigators and were grouped according to four broad categories: population discussed (the subject of the social media post; eg, "women of reproductive age," "parents," and "pediatric patients"); awareness and knowledge (a list of clinical and nonclinical terms related to CMV; eg, "seroconversion," "newborn screening," "National CMV Foundation," and "parental education"); prevention (a list of terms describing preventive

measures related to CMV; eg, "hygiene measures," "antiviral treatment," and "vaccines"); and general CMV information (generic terms; eg, "safety," "efficacy," and "tolerability"). In addition, the perceived target audience (eg, general population, scientists/health care professionals) was annotated (see Table S1 in Multimedia Appendix 3 for a complete list of prespecified aspects within each thematic category). Investigators manually classified prespecified aspects for approximately 90 social media posts. The ChatGPT model then used the manually classified social media posts for few-shot learning for aspect classification. Although 90 manually classified posts represented a small subset of the >10,000 posts analyzed, the approach ensured complete coverage of all prespecified aspects with representation in at least 2 example posts, making the number of manually classified posts commensurate with the list of aspects of interest (Table S1 in Multimedia Appendix 3).

To load the full dataset stored in a Microsoft Excel spreadsheet, the tweets were imported into data frames using the Python *pandas* library, and only the tweet-text column was used as input for annotation. The master prompt (Multimedia Appendix 2) contained (1) step-by-step instructions for identifying CMV-related aspects, segmenting text, and assigning sentiment, while maintaining aspect integrity, and (2) the set of approximately 90 manually annotated tweets used as few-shot examples. The model temperature was set to 0 to promote deterministic outputs. Responses were requested in JSON format (keys included aspects, aspect segment, sentiment toward each aspect, and overall sentiment; see sentiment methods in the *Sentiment and other Statistical Analyses* section). Each tweet was processed with up to five retries to ensure valid JSON formatting. No manual prompt revisions were made between batches, and no additional model fine-tuning was performed. Reproducibility was supported by fixing model parameters (temperature=0), using a single master prompt for all tweets, processing tweets independently, and allowing up to five retries for valid JSON output. To enhance transparency, the prompt instructed ChatGPT to include additional keys in the JSON output, identifying the specific text segments that contributed to each aspect or sentiment label. This allowed investigators to trace which words or phrases informed each annotation. Because annotations were produced via prompt-based generation rather than model training, no explicit class-balancing techniques were applied.

Sentiment and Other Statistical Analyses

Social media posts and specific aspects identified within the posts were assigned a sentiment (ie, positive, neutral, or negative). Four independent, blinded reviewers assigned a sentiment to 97 randomly selected post texts outside of the dataset with moderate agreement, with an interrater reliability score of 0.56 (Fleiss κ; "moderate" defined as 0.41-0.6 [42]). The ChatGPT model was then provided with the scored posts for few-shot learning and asked to assign a sentiment to social media posts included in the analysis dataset. Following assignment, text from 50 posts scored by ChatGPT was evaluated for sentiment by the same four independent and blinded reviewers. Moderate interrater reliability was measured at 0.51 (Fleiss κ). The interrater reliability score between the sentiment assigned by most reviewers and ChatGPT was 0.69 (Cohen κ), indicating substantial agreement between reviewers and ChatGPT ("substantial" defined as 0.61-0.8 [42]). Of the 10 posts, or 20%, where human-AI agreement was not observed, 90% of posts were scored neutral by one party and either positive or negative by the other, indicating that minor discrepancies rather than large errors (positive versus negative) explain the discordance.

Statistical analyses (Fleiss κ, Cohen κ, chi-square tests with Bonferroni correction, Kruskal-Wallis with Bonferroni correction) were conducted using Excel or Datatab. Chi-square tests were applied only to independent categorical variables. Individual posts often contained multiple dependent variables (eg, multiple hashtags; multiple aspects within a category, such as target audience, awareness, or prevention messaging). These attributes were excluded from tests requiring variable independence. Adjusted Pearson residuals were calculated for chi-square cross-tabulations, and Bonferroni correction was applied to adjust *P* values for multiple comparisons to set the critical threshold for significance.

Ethical Considerations

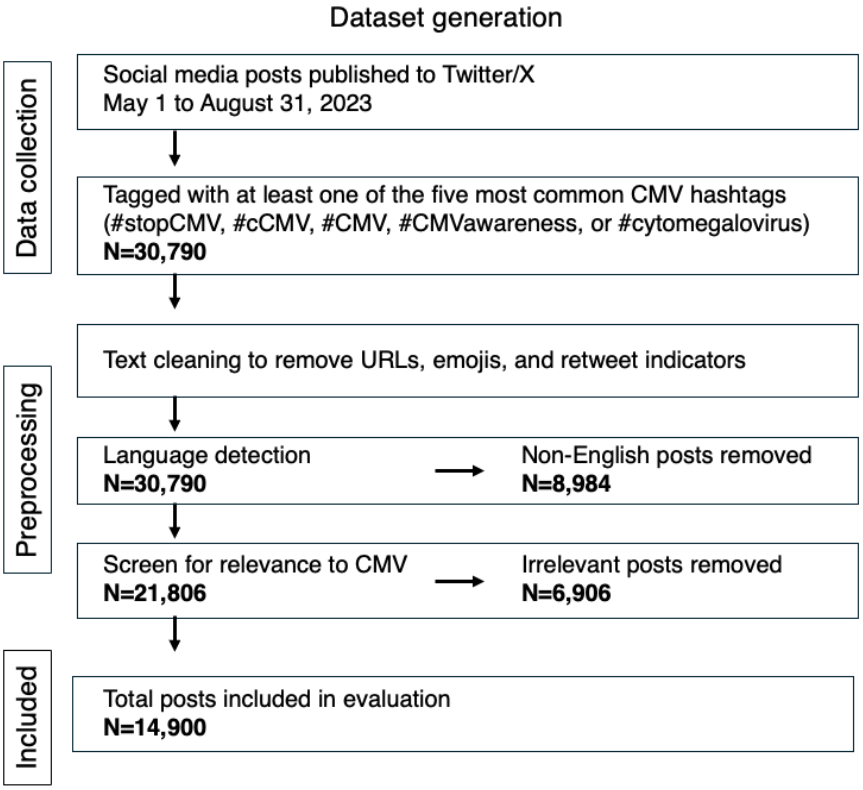
This study used only publicly available text from social media posts on Twitter/X. The research team did not interact with post authors or collect private information. All data were processed to remove or avoid inclusion of identifiable information and are reported only in aggregate. This study did not constitute human subjects research and therefore did not require review or approval by an institutional review board per Federal Regulations for the Protection of Human Research Subjects (45 CFR 46.104(d)(4)(i); [43]).

Results

Data Aggregation and Processing

Between May 1 and August 31, 2023, a total of 30,790 public posts published to Twitter/X were tagged with one or more of the five most common hashtags used to reference CMV or CMV disease. Several hundred posts were cotagged with hashtags unrelated to CMV (eg, #SHREKINU, a cryptocurrency) or used a CMV-related hashtag to indicate a non-CMV topic (eg, #CMV=commercial motor vehicle). Additionally, nearly one-third of posts were written in a language other than English. To focus subsequent analyses, the 30,790 posts were preprocessed to (1) remove noise elements, such as URLs and emojis; (2) filter out non-English posts; and (3) extract posts relevant to CMV or CMV disease (Figure 1).

Figure 1. Flowchart of the dataset collection, preprocessing, and inclusion steps. cCMV: congenital cytomegalovirus; CMV: cytomegalovirus.



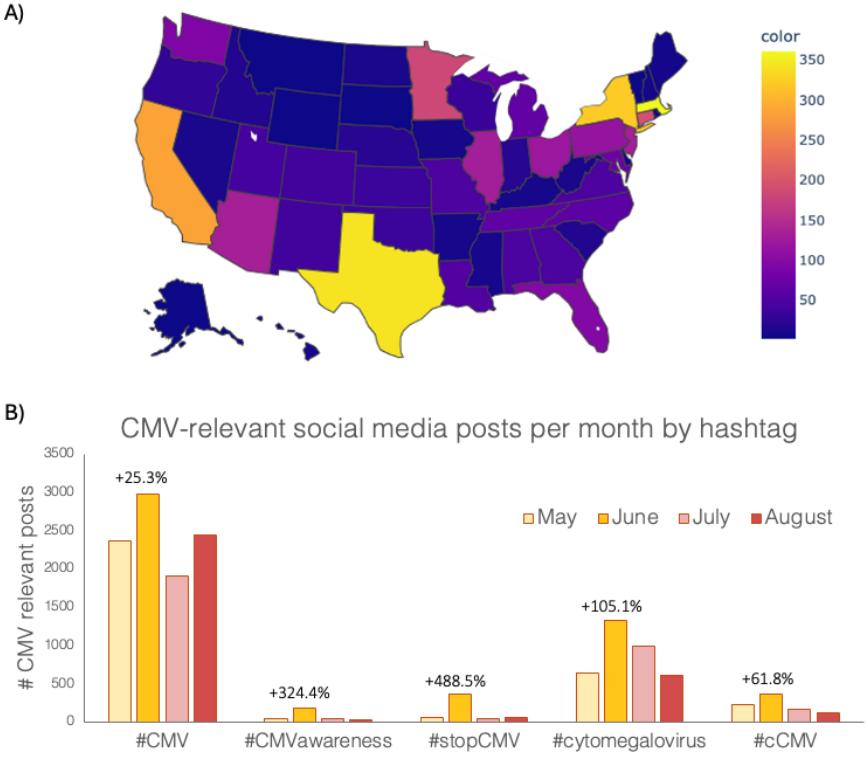
Descriptive Analyses

Overall, 14,900 posts were analyzed. There were 3336 (22.4%) social media posts in May 2023, which increased by 55% to 5180 (34.8%) posts in June 2023 (coinciding with the National CMV Awareness Month in the United States), and decreased during both July 2023 (n=3124, 21%) and August 2023 (n=3260, 21.9%), as shown in Table S2 in Multimedia Appendix 3. Social media posts most frequently originated from the United States (n=3858, 25.9%). The next most frequent countries included Canada (n=594, 4%), the United Kingdom (n=535, 3.6%), India (n=384, 2.6%), and Australia (n=370, 2.5%). In the United States, among posts with known geographic locations, Massachusetts, Texas, and New York had the highest number of posts (Figure 2A). A more detailed analysis of posts grouped by hashtag revealed additional trends. The hashtag #CMV was the most used (n=9674, 64.9%), followed by #cytomegalovirus (n=3558, 23.9%), #cCMV (n=859, 5.8%), #stopCMV (n=529, 3.6%), and #CMVawareness (n=280, 1.9%), as shown in Figure 2B. With respect to the National CMV Awareness Month, an increase in posts was observed from May to June for all five hashtags. The largest proportional increases occurred with #stopCMV (488.5%, from n=61, 0.4%, to n=359, 2.4%, posts) and #CMVawareness (+324.4%; from n=41, 0.3%, to n=174, 1.2%, posts), possibly driven by national advocacy organizations.

To understand who is publishing CMV-relevant content on social media, the authors of posts containing each CMV-related hashtag were analyzed. Because a single post may include multiple hashtags, authors could appear across several hashtag groups. For each hashtag, authors were ranked by the total number of posts they published, and the top 20 most frequent authors were identified. These authors were then categorized by author category based on their profile affiliation. To describe overall trends, the top 20 authors from each hashtag were combined into a single dataset, with duplicate users removed. The most common author category was “university/academic researchers,” followed by “news/journal/public health/education and “physicians/health care/hospitals” (Table S3 in Multimedia Appendix 3).

Hashtag-specific trends included the nearly equal representation of most CMV stakeholders under #CMV, in contrast to skewed author distributions for #CMVawareness and #cytomegalovirus. The top 20 authors scored by total followers were also assessed. Unsurprisingly, the most common follower category was “news/journal/public health/education” (Table S3 in Multimedia Appendix 3).

Figure 2. Descriptive statistics. Quantification of CMV-relevant public posts (A) published to Twitter/X at the US state level and (B) sorted by hashtag by month (May-August 2023). Noted percentages reference the change in post number from May to June (National CMV Awareness Month). cCMV: congenital cytomegalovirus; CMV: cytomegalovirus.

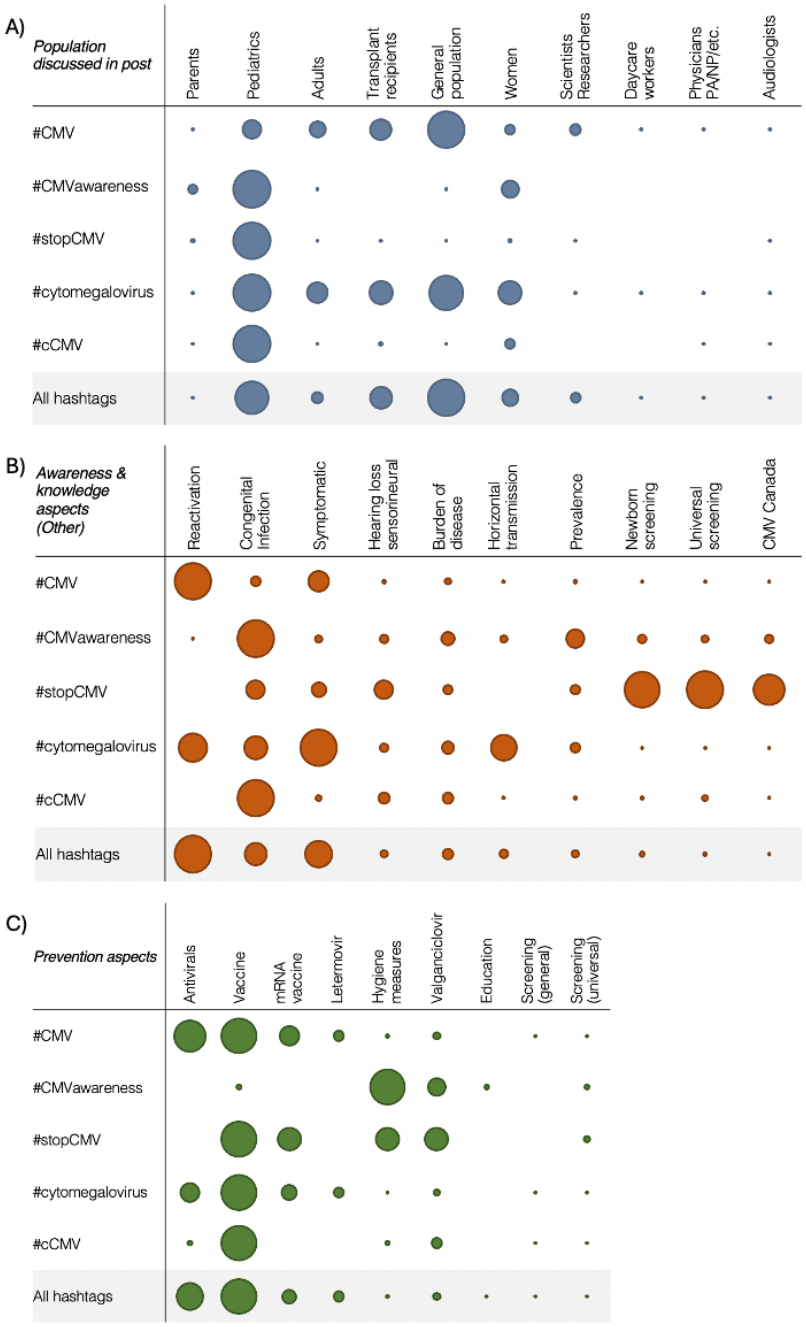


Classification of Thematic (Aspects) Content

Social media posts were also evaluated for thematic content. Five broad thematic categories (target audience, population discussed, awareness and knowledge, prevention, and general CMV information) were created and populated with specific aspects to be identified within the posts. Overall, the "scientific/health care professionals" category was the most frequent target audience of social media posts (n=7575, 50.8%); however, this varied by hashtag (Table S4 in Multimedia Appendix 3). The "general population" category was the most frequent population discussed across all five hashtags (n=2727, 18.3%), followed by "pediatrics" (n=2456, 16.5%), as shown in Figure 3A and in Table S5 in Multimedia Appendix 3. Transplant recipients (n=1636, 11%) were the third-most frequent population discussed, though they were nearly exclusively mentioned in #CMV and #cytomegalovirus posts. "Women" (n=1141, 7.7%), "adults" (n=784, 5.3%), and "scientists" (n=691, 4.6%) were also frequently discussed; all other prespecified aspects related to the "population discussed" category were identified in fewer than 2235 (15%) of the 14,900 social media posts evaluated.

Approximately 50 aspects were prespecified for the "awareness and knowledge" category (Table S1 in Multimedia Appendix 3). Perhaps expectedly, "CMV" was overwhelmingly the most frequently identified aspect (n=10,929, 73.3%, posts), with "cCMV" being the second-most frequently identified aspect (n=1598, 10.7%), as shown in Table S6 in Multimedia Appendix 3. The frequency of aspects other than "CMV" and "cCMV" varied by hashtag, with the third-most frequent aspect being "reactivation" for #CMV and #cytomegalovirus, "congenial infection" for #CMVawareness and #cCMV, and "newborn/universal screening" for #stopCMV (Table S7 in Multimedia Appendix 3 and Figure 3B). Although "prevention" aspects were included in just 2593 (17.4%) of the 14,900 posts, the most frequent preventions identified included "vaccines" (n=1115, 7.5%), "antiviral treatment" (n=849, 5.7%), "mRNA vaccine" (n=453, 3%), and "Letermovir" (n=332, 2.2%), as shown in Figure 3C and in Table S8 in Multimedia Appendix 3. "Hygiene measures" and the specific antiviral treatment "Valganciclovir" were less common overall (n=114, 0.8%, and n=214, 1.4%, respectively), though these were the most frequent "prevention" aspects within #CMVawareness posts. In the final category, which included terms reflecting "general CMV information," "immune response" was the most frequently identified aspect (n=533, 3.6%), which was two times greater than "side effects," the second-most frequently identified aspect in this category (Table S9 in Multimedia Appendix 3). In general, aggregate results combining data for all five hashtags were consistent with results for the individual hashtags #CMV and #cytomegalovirus; however, as described here, distinct trends regarding thematic content were observed when quantified by individual hashtags.

Figure 3. Relative comparison of the number of posts by aspect for each hashtag or aggregated. (A) Population discussed in a post, (B) awareness aspects other than CMV or cCMV, and (C) CMV prevention aspects. The largest circle in a row represents the highest number of posts with that aspect for that hashtag. All other circles are proportional to this reference. cCMV: congenital cytomegalovirus; CMV: cytomegalovirus; mRNA: messenger RNA.



The Impact of the National CMV Awareness Month on Thematic Content of Posts

We next assessed the effect of the National CMV Awareness Month (June 2023) on the thematic content of Twitter/X posts. In our sample of 12,910 (86.6%) posts from May to July 2023, a significant association was observed between the month of publication and the target audience (χ^2_2=144.3, *P*<.001), with an above-expected increase in the general population in June and a decrease in scientists/health care professionals (adjusted Pearson residuals; see Table S10 in Multimedia Appendix 3). This shift from scientists/health care professionals to the general population from May to June was lost in July (Tables S11-S13 in Multimedia Appendix 3). The population discussed also shifted during the National CMV Awareness Month, with a 110% increase in the "pediatrics" group (n=577, 4.5%, posts in May to n=1213, 9.4%, in June), a 134% increase in "women" (n=280, 2.2%, posts in May to n=656, 5.1%, in June), and a decrease of –32% in "transplant recipients" (n=569, 4.4%, posts in May to n=386, 3%, in June), as shown in Tables S11-S13 in Multimedia Appendix 3. Given the goal of the National CMV Awareness Month to bring awareness about CMV and prevent infection, we next tested for an association between the month of publication and the number of posts that contained these attributes. A significant association was observed between "awareness" and "prevention" and the publication month

(χ^2_2=107.8, *P*<.001), with a significant increase in posts containing awareness messaging compared to that expected and a significant decrease in prevention messaging in June (adjusted Pearson residuals; see Table S10 in Multimedia Appendix 3). To understand which awareness and prevention aspects shifted during the National CMV Awareness Month, we analyzed individual aspects. Within the "awareness and knowledge" thematic category, the absolute number of posts increased across all aspect groups from May to June 2023 (Tables S11-S13 in Multimedia Appendix 3). We then assessed whether the proportional representation of individual aspects also shifted during the National CMV Awareness Month. The proportion of "awareness and knowledge" posts mentioning "burden of disease" increased from 5.7% (59/3773) of posts in May to 9.8% (173/6421) in June (+71%), and posts mentioning "horizontal transmission" increased from 3.8% (39/3773) of posts in May to 10.4% (183/6421) in June (+174%). In contrast, mentions of "universal screening" decreased from 5.6% (58/3773) of posts in May to 3.6% (64/6421) in June (–35%), as shown in Tables S11-S13 in Multimedia Appendix 3. Shifts in the proportional representation of aspects within the "prevention" category also occurred from May to June 2023. The proportion of "prevention" posts with mentions of "vaccine" increased from 24.5% (207/845) of posts in May to 42.9% (368/857) in June (+75%). A proportional decrease in posts that mentioned "antivirals" (–26%; 268/845, 31.7%, posts in May to 202/857, 23.6%, posts in June) and "hygiene measures" (–37%; 50/845, 5.9%, posts in May to 32/857, 3.7%, posts in June) also occurred (Tables S11-S13 in Multimedia Appendix 3). Finally, as expected, the "fundraising" aspect increased within the "general" category of posts during the National CMV Awareness Month (+157%; 37/286, 4.4%, posts in May to 95/329, 11.2%, posts in June), as shown in Tables S11-S13 in Multimedia Appendix 3.

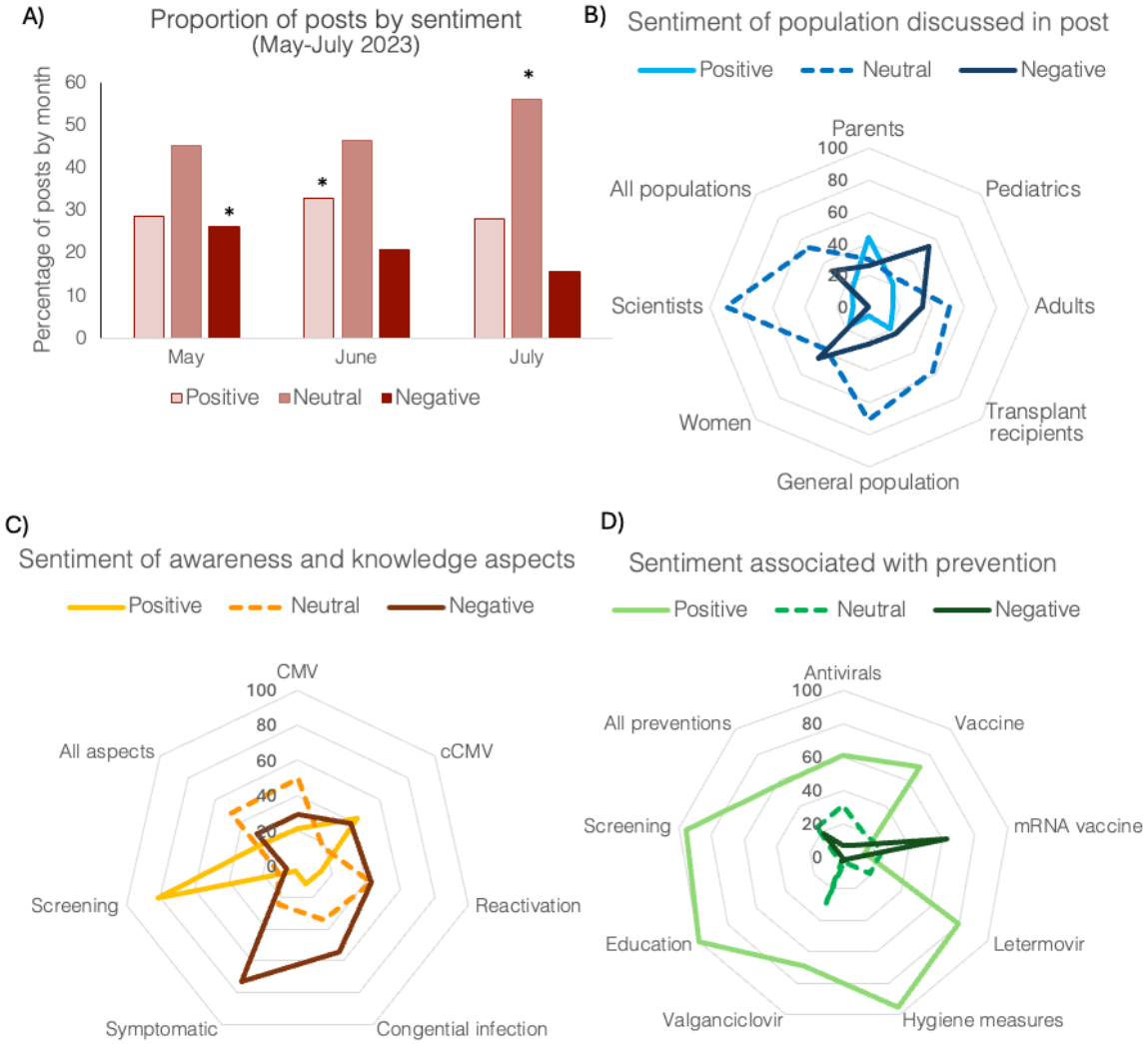
Sentiment Analyses

Through sentiment analysis, the majority of social media posts were classified as neutral. This was true for most of the broad thematic categories of prespecified aspects (Table S14 in Multimedia Appendix 3 and Figure 4A), though the trend reversed for aspects pertaining to "prevention" for which the number of posts classified as positive (n=1288, 8.6%) was several times greater than the number of posts classified as negative (n=418, 2.8%), as shown in Table S17 in Multimedia Appendix 3. Overall, more social media posts were classified as positive relative to negative (n=3531, 23.7%, vs n=2436, 16.3%, respectively; see Table S14 in Multimedia Appendix 3). In our sample of 12,910 (86.6%) posts from May to July 2023, a significant association was observed between sentiment type and the month of publication (χ^2_4=163.6, *P*<.001). An examination of all possible combinations of posts revealed posts during the National CMV Awareness Month to be significantly more likely to be positive than expected (adjusted Pearson residuals; see Table S14 in Multimedia Appendix 3 and Figure 4A).

The sentiment associated with independent aspects was also scored. A significant association was observed between sentiment type and the specific target audience, either the general population or scientists/health care professionals (χ^2_2= 481.2, *P*<.001). Although most posts were classified as neutral (Table S15 in Multimedia Appendix 3), "general population" posts were significantly more likely to be positive or negative in sentiment than expected, while "scientists/health care professionals" posts were significantly more likely to be neutral (adjusted Pearson residuals; see Table S15 in Multimedia Appendix 3).

The sentiment varied depending on the specific "population" discussed," the specific "awareness and knowledge" aspect, or the type of "prevention." For example, among the populations discussed, the sentiment of the "scientists" aspect was disproportionally neutral (620/691, 89.7%); in contrast, the sentiment of the "parents" aspect was more likely classified as positive (99/225, 44%), and the most common sentiment for the "pediatrics" aspect was negative (1336/2476, 54%), as shown in Table S16 in Multimedia Appendix 3 and Figure 4B. Examples of aspects included in the "awareness and knowledge" category that diverged from the average sentiment included increased positive association with "screening" (266/327, 81.3%) and increased negative association with "symptomatic" (712/977, 72.9%) and "congenital infection" (437/800, 54.6%), as shown in Table S17 in Multimedia Appendix 3 and Figure 4C. Lastly, "education," "screening," and "hygiene measures" aspects included in the "prevention" category were scored as nearly universally positive (education: 23/23, 100%;screening: 22/23, 95.7%; and hygiene measures: 109/114, 95.6%), as shown in Table S18 in Multimedia Appendix 3 and Figure 4D.

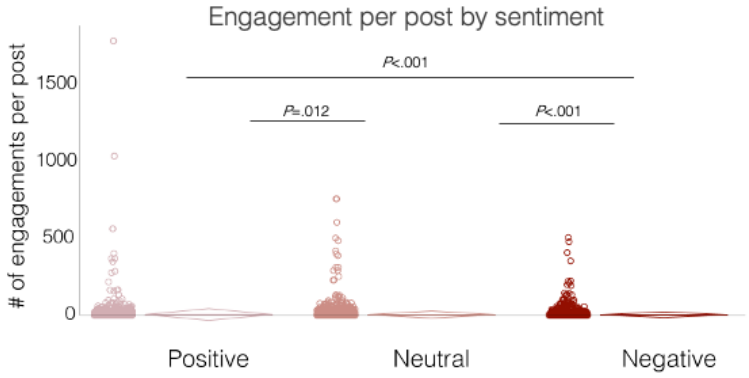
Figure 4. Evaluation of post sentiment from May to July 2023. (a) Proportion of posts scored by overall sentiment per month, (b) by population discussed, (c) by awareness and knowledge, and (d) by type of prevention. Post attributes that occurred infrequently (<10% of the total aspects) were not included. Posts that included more than one attribute are represented in each individual aspect that the posts included. *Significant increase over expected (*P*<.05), chi-square, adjusted Pearson residuals. cCMV: congenital cytomegalovirus; CMV: cytomegalovirus; mRNA: messenger RNA.



User Engagement

In our final analysis, we assessed user engagement with CMV-relevant posts, including "likes," retweets, and comments. Overall, users engaged with 3863 (27.3%) of 14,136 posts from May to August 2023. Post engagement was slightly higher on average if scored with a positive sentiment (1230/3895, 31.6%) compared to a neutral (1923/6445, 29.8%) or a negative (710/3796, 18.7%) sentiment (Table S19 in Multimedia Appendix 3). To understand potential differences in the magnitude of engagement with respect to sentiment, we compared median engagement per post across sentiment groupings. Engagement was found to be significantly different between sentiments (Kruskal-Wallis; χ^2_2=194.31, *P*<.001), with a higher rank mean for positive posts compared to neutral and negative posts (Table S19 in Multimedia Appendix 3 and Figure 5). Therefore, posts with a positive sentiment were more likely to engage the audience and to a greater degree, while those with negative sentiment were least likely to do so.

Figure 5. Evaluation of post engagement with respect to sentiment (May-August 2023). A dot plot of the number of engagements per post categorized by sentiment, with the SD displayed to the right. **P* values are adjusted with Bonferroni correction, Kruskal-Wallis, Dunn-Bonferroni test. One outlier value (3835, negative) was removed from the graph to improve visualization of data.



Discussion

Principal Findings

The objective of our study was to monitor the volume and thematic content of social media posts on Twitter/X before, during, and after the National CMV Awareness Month in June 2023 to understand the virtual impact of the campaign. We first used a language detection model and a customized BERT tokenizer to extract English language tweets and to remove posts that were not relevant to CMV or CMV disease. Analysis of the remaining 14,900 CMV-relevant posts revealed a peak in posts during the National CMV Awareness Month, a trend observed across all five of the most frequently used CMV-related hashtags, with the highest volume of posts originating from the United States. As expected, these data confirm an active campaign initiative by multiple CMV stakeholders in the United States. We sought to further characterize who is generating information and how effectively their messages are disseminated by identifying the key users and classifying these authors by affiliation. These data point to a potential opportunity to enhance collaboration between advocacy organizations, academic researchers (who were observed to be the most prolific authors), and media outlets (observed to have the largest number of followers) to expand messaging and to target messaging to specific audiences.

Our analyses also examined the thematic content, or aspects, of social media posts, along with the sentiment of each post. A subset of posts (outside of the primary dataset) was annotated by researchers for these attributes, and this dataset was provided to a ChatGPT model, which annotated CMV-relevant posts in substantial agreement with blinded reviewers. Highly mentioned aspects were typically shared between hashtags, though hashtag-specific trends point to specific conversations being siloed to separate channels. For example, “screening” was mentioned primarily under the #stopCMV hashtag, while “transplant recipients” were discussed under the #CMV and #cytomegalovirus hashtags. The National CMV Awareness Month shifted CMV conversations toward the general audience from scientists and health care professionals and were more likely to contain awareness messaging indicating an effort by stakeholders to increase attention with respect to pediatric populations, women, and the burden of disease, most likely reflecting advocacy around cCMV. The sentiment of CMV-relevant social media posts was overall neutral, though it shifted meaningfully toward positive during the campaign month. The overwhelmingly positive sentiment associated with CMV the “prevention” aspects “education,” “screening,” and “hygiene measures” speaks to the enthusiastic advocacy of the CMV community for various interventions and preventive measures. Unexpectedly, prevention messaging decreased significantly during the awareness month even as posts containing prevention aspects were observed to have a positive sentiment, which correlates with higher community engagement.

The methodology used in this study was implemented with multiple checks to ensure accuracy in processing and analysis and is described in sufficient detail to support reproducibility. This allows our methodology to be easily used to evaluate future CMV awareness campaigns to identify long-term trends or shifts in CMV thematic content and sentiment. We envision the possibility of monitoring social media posts as new interventions become available, as is done commonly for sentiment analysis around vaccines against influenza and SARS-CoV-2. Our methods are adaptable and can be expanded to monitor messaging once a vaccine against CMV becomes available or as new legislation is proposed around newborn screening for cCMV. These surveillance and data collection steps provide a foundation for more rigorous interpretation through the lens of health communication and health behavior theory to inform future advocacy and awareness campaigns.

Limitations

This study has several limitations. We examined posts from a single social media platform (ie, Twitter/X) using Keyhole. Users of Twitter/X are not necessarily representative of the broader US population or all CMV stakeholders; moreover, user demographics likely differ on other platforms, such as Instagram, Reddit, Facebook, and TikTok. Furthermore, although Keyhole is widely used in commercial practice, it is not commonly used in public health studies, in part due to cost. The platform aligns well with real-world social listening practice (eg, real-time tracking), making it suitable for evaluating a public health awareness campaign; however, data were collected, filtered, and summarized by Keyhole rather than accessed as raw posts directly from the Twitter/X API, which may affect completeness and reproducibility.

We were also unable to analyze nearly 9000 non-English tweets, which may differ in terms of author category, aspects, and sentiment relative to the English-language posts analyzed here. Furthermore, the manual classification of aspects and sentiment by researchers is a subjective process, and the list of aspects generated by the authors, while detailed and broad, may not reflect all possible themes discussed in posts. Although moderate in agreement, the interrater reliability in sentiment scores between the four independent raters reflects a limitation in the sentiment analysis. Complete disagreement between the reviewers occurred in 4 of 50 test tweets (positive or negative, excluding neutral), which may reflect the inherently mixed nature of disease-related posts that combine positive messaging around interventions or milestones and negative messaging that describes the underlying need for the intervention. Accurate annotation and classification by AI were dependent on these subjective processes. They may have also been impacted by the constrained length of tweets, which can limit content, explanatory details, and other cues. Lastly, although the pre- and postperiods surrounding June 2023 we examined are limited, we were able to evaluate the immediate impact of the National CMV Awareness Month. For these reasons, the results reported here should be viewed as exploratory and interpreted with this lens. Additional research leveraging the expertise of diverse stakeholders is needed to design and evaluate long-term public health information and future CMV awareness campaigns.

Strengths

This study also has several notable strengths. To the best of our knowledge, it represents the first systematic infodemiologic evaluation of the National CMV Awareness Month and the first large-scale characterization of conversations concerning CMV on Twitter/X. By analyzing nearly 15,000 CMV-relevant posts, this study leveraged a larger and more comprehensive dataset than would have been feasible to analyze through manually annotation alone. Furthermore, analysis from multiple lenses, such as user characteristics, thematic content, and sentiment, provides a multidimensional view of CMV discussions that has not been unavailable to researchers, advocates, or other relevant stakeholders.

A second strength is the practical demonstration of a multistep analytical pipeline that combines human annotation, model customization, and iterative validation. Furthermore, the application of few-shot prompting with ChatGPT enabled efficient classification of aspects and sentiment at scale. Importantly, human-AI agreement of post sentiment was substantial, increasing confidence in the reliability of the automated post annotations. Finally, the study's design, which evaluated posts before, during, and after the campaign month, allows for a direct observation of how a nationally recognized campaign shifts online attention, thematic content, and sentiment.

Conclusion

To the best of our knowledge, this study is the first to examine and report the volume, thematic content, and sentiment of virtual CMV-related conversations on Twitter/X before, during, and after the National CMV Awareness Month. The use of AI permitted detailed evaluation of many thousands of social media posts. The results of our analyses enable us to predict potential collaborations between key users to achieve greater dissemination and impact during future campaigns. In addition, the detailed analyses presented here provide a more complete characterization of the conversations and culture within distinct CMV-related hashtags and highlight the thematic content that can be amplified in future campaigns. The complexity of health communication via social media poses distinct challenges to public health investigators and practitioners when planning and executing information and awareness campaigns. Although this study demonstrates the analytic capabilities of AI, the generative capabilities of the ChatGPT model could also be used to draft campaign messaging to enhance specific themes or emotional undertones.

Acknowledgments

We are grateful to Giovana Vieira and Rofail Wassef for their kind contribution of scoring the sentiment for social media posts. This research was supported by startup funds provided to TR.

Data Availability

The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization was handled by KF and TRR; data curation, ZY and TRR; formal analysis, ZY, TRR, LD, and TP; funding acquisition, TRR; investigation, TRR, ZY, LD, TP, and KF; methodology, ZY, TP, TRR, JDD, and RVW; project administration, CK, TRR, JDD, and RVW; resources, TRR; supervision, CK, KF, TRR, and JDD; validation, TRR and ZY; visualization, TRR and ZY; writing—original draft, TRR, CK, ZY, and LD; and writing—review and editing, TRR, CK, ZY, LD, RVW, KF, and JDD. No advocacy organization provided funding for this study. The views expressed are those of the authors and do not necessarily represent those of their institutions.

Conflicts of Interest

TRR is an employee of Stonehill College, and LD was an undergraduate student. TRR received institutional startup funds from Stonehill College that supported this work. ZY, TP, and RVW were employees of Moderna Therapeutics, Incorporated, at the

time this study was conducted and may have held company stock or stock options during that period. JDD and CK are employees of Moderna Therapeutics, Incorporated, and may hold company stock or stock options. KF receives consulting fees from Moderna Therapeutics and KF reports no competing interests. The authors have no financial relationships with Twitter/X, Keyhole, OpenAI, or the National Cytomegalovirus Foundation relevant to this work.

Multimedia Appendix 1
Expanded methodology.
[DOCX File , 20 KB - infodemiology_v6i1e80922_app1.docx]

Multimedia Appendix 2
ChatGPT-4 model master prompt used to annotate aspects and sentiments and identify tweet segments associated with those aspects and sentiments.
[DOCX File , 17 KB - infodemiology_v6i1e80922_app2.docx]

Multimedia Appendix 3
Thematic categories, social media post counts, top 20 social media authors, adjusted Pearson residuals and chi-square test results, and engagement with respect to sentiment.
[DOCX File , 61 KB - infodemiology_v6i1e80922_app3.docx]

References

1. Dana Flanders W, Lally C, Dilley A, Diaz-Decaro J. Estimated cytomegalovirus seroprevalence in the general population of the United States and Canada. J Med Virol 2024 Mar 26;96(3):e29525 [FREE Full text] [doi: 10.1002/jmv.29525] [Medline: 38529529]
2. Lantos PM, Permar SR, Hoffman K, Swamy GK. The excess burden of cytomegalovirus in African American communities: a geospatial analysis. Open Forum Infect Dis 2015 Dec;2(4):ofv180 [FREE Full text] [doi: 10.1093/ofid/ofv180] [Medline: 26716106]
3. Zuhair M, Smit GSA, Wallis G, Jabbar F, Smith C, Devleesschauwer B, et al. Estimation of the worldwide seroprevalence of cytomegalovirus: a systematic review and meta-analysis. Rev Med Virol 2019 May 31;29(3):e2034 [FREE Full text] [doi: 10.1002/rmv.2034] [Medline: 30706584]
4. Lantos PM, Hoffman K, Permar SR, Jackson P, Hughes BL, Kind A, et al. Neighborhood disadvantage is associated with high cytomegalovirus seroprevalence in pregnancy. J Racial Ethn Health Disparities 2018 Aug;5(4):782-786 [FREE Full text] [doi: 10.1007/s40615-017-0423-4] [Medline: 28840519]
5. Clinical overview of CMV and congenital CMV: cytomegalovirus (CMV) and congenital CMV infection. Centers for Disease Control and Prevention. 2024. URL: https://www.cdc.gov/cytomegalovirus/hcp/clinical-overview/index.html [accessed 2024-12-16]
6. Stagno S, Pass RF, Dworsky ME, Alford CA. Maternal cytomegalovirus infection and perinatal transmission. Clin Obstet Gynecol 1982 Sep;25(3):563-576 [FREE Full text] [doi: 10.1097/00003081-198209000-00014] [Medline: 6290121]
7. Wang H, Peng G, Bai J, He B, Huang K, Hu X, et al. Cytomegalovirus infection and relative risk of cardiovascular disease (ischemic heart disease, stroke, and cardiovascular death): a meta‐analysis of prospective studies up to 2016. J Am Heart Assoc 2017 Jul 06;6(7):e005025 [FREE Full text] [doi: 10.1161/JAHA.116.005025] [Medline: 28684641]
8. Aiello AE, Haan M, Blythe L, Moore K, Gonzalez JM, Jagust W. The influence of latent viral infection on rate of cognitive decline over 4 years. J Am Geriatr Soc 2006 Jul;54(7):1046-1054 [FREE Full text] [doi: 10.1111/j.1532-5415.2006.00796.x] [Medline: 16866674]
9. Burgdorf KS, Trabjerg BB, Pedersen MG, Nissen J, Banasik K, Pedersen OB, et al. Large-scale study of toxoplasma and cytomegalovirus shows an association between infection and serious psychiatric disorders. Brain Behav Immun 2019 Jul;79:152-158 [FREE Full text] [doi: 10.1016/j.bbi.2019.01.026] [Medline: 30685531]
10. Barnes LL, Capuano AW, Aiello AE, Turner AD, Yolken RH, Torrey EF, et al. Cytomegalovirus infection and risk of Alzheimer disease in older black and white individuals. J Infect Dis 2015 Jan 15;211(2):230-237 [FREE Full text] [doi: 10.1093/infdis/jiu437] [Medline: 25108028]
11. Phillips AC, Carroll D, Khan N, Moss P. Cytomegalovirus is associated with depression and anxiety in older adults. Brain Behav Immun 2008 Jan;22(1):52-55 [FREE Full text] [doi: 10.1016/j.bbi.2007.06.012] [Medline: 17703915]
12. Gadoth A, Ourfalian K, Basnet S, Kunzweiler C, Bohn RL, Fülöp T, et al. Potential relationship between cytomegalovirus and immunosenescence: evidence from observational studies. Rev Med Virol 2024 Jul;34(4):e2560 [FREE Full text] [doi: 10.1002/rmv.2560] [Medline: 38866595]
13. Orlikowski D, Porcher R, Sivadon-Tardy V, Quincampoix J, Raphaël JC, Durand M, et al. Guillain-Barré syndrome following primary cytomegalovirus infection: a prospective cohort study. Clin Infect Dis 2011 Apr 01;52(7):837-844 [FREE Full text] [doi: 10.1093/cid/cir074] [Medline: 21427390]

14. Mercado NB, Real JN, Kaiserman J, Panagioti E, Cook CH, Lawler SE. Clinical implications of cytomegalovirus in glioblastoma progression and therapy. NPJ Precis Oncol 2024 Sep 29;8(1):213 [FREE Full text] [doi: 10.1038/s41698-024-00709-4] [Medline: 39343770]
15. Savva GM, Pachnio A, Kaul B, Morgan K, Huppert FA, Brayne C, Medical Research Council Cognitive Function and Ageing Study. Cytomegalovirus infection is associated with increased mortality in the older population. Aging Cell 2013 Jun;12(3):381-387 [FREE Full text] [doi: 10.1111/acel.12059] [Medline: 23442093]
16. Leeaphorn N, Garg N, Thamcharoen N, Khankin EV, Cardarelli F, Pavlakis M. Cytomegalovirus mismatch still negatively affects patient and graft survival in the era of routine prophylactic and preemptive therapy: a paired kidney analysis. Am J Transplant 2019 Feb;19(2):573-584 [FREE Full text] [doi: 10.1111/ajt.15183] [Medline: 30431703]
17. Stern L, Withers B, Avdic S, Gottlieb D, Abendroth A, Blyth E, et al. Human cytomegalovirus latency and reactivation in allogeneic hematopoietic stem cell transplant recipients. Front Microbiol 2019 May 28;10:1186 [FREE Full text] [doi: 10.3389/fmicb.2019.01186] [Medline: 31191499]
18. CMV fact sheet for healthcare providers. Centers for Disease Control and Prevention. URL: https://stacks.cdc.gov/view/cdc/137322 [accessed 2025-12-23]
19. Boppana SB, van Boven M, Britt WJ, Gantt S, Griffiths PD, Grosse SD, et al. Vaccine value profile for cytomegalovirus. Vaccine 2023 Nov 03;41 Suppl 2:S53-S75 [FREE Full text] [doi: 10.1016/j.vaccine.2023.06.020] [Medline: 37806805]
20. Lombardi G, Garofoli F, Stronati M. Congenital cytomegalovirus infection: treatment, sequelae and follow-up. J Matern Fetal Neonatal Med 2010 Oct;23 Suppl 3:45-48 [FREE Full text] [doi: 10.3109/14767058.2010.506753] [Medline: 20807160]
21. Song X, Li Q, Diao J, Li J, Li Y, Zhang S, et al. Association between first-trimester maternal cytomegalovirus infection and stillbirth: a prospective cohort study. Front Pediatr 2022 Jun;10(11):803568-803604 [FREE Full text] [doi: 10.3389/fped.2022.803568] [Medline: 35372174]
22. Pesch M, Mowers J, Huynh A, Schleiss M. Intrauterine fetal demise, spontaneous abortion and congenital cytomegalovirus: a systematic review of the incidence and histopathologic features. Viruses 2024 Sep 30;16(10):1552 [FREE Full text] [doi: 10.3390/v16101552] [Medline: 39459885]
23. Pereira L, Petitt M, Fong A, Tsuge M, Tabata T, Fang-Hoover J, et al. Intrauterine growth restriction caused by underlying congenital cytomegalovirus infection. J Infect Dis 2014 May 15;209(10):1573-1584 [FREE Full text] [doi: 10.1093/infdis/jiu019] [Medline: 24403553]
24. Manicklal S, Emery VC, Lazzarotto T, Boppana SB, Gupta RK. The “silent” global burden of congenital cytomegalovirus. Clin Microbiol Rev 2013 Jan;26(1):86-102 [FREE Full text] [doi: 10.1128/cmr.00062-12]
25. Diaz-Decaro J, Myers E, Mucha J, Neumann M, Lewandowski W, Kaczanowska M, et al. A systematic literature review on the humanistic burden of cytomegalovirus. Curr Med Res Opin 2023 May;39(5):739-750 [FREE Full text] [doi: 10.1080/03007995.2023.2191477] [Medline: 36938652]
26. Diaz-Decaro J, Myers E, Mucha J, Neumann M, Lewandowski W, Kaczanowska M, et al. A systematic literature review of the economic and healthcare resource burden of cytomegalovirus. Curr Med Res Opin 2023 Jul;39(7):973-986 [FREE Full text] [doi: 10.1080/03007995.2023.2222583] [Medline: 37395088]
27. Cannon MJ, Westbrook K, Levis D, Schleiss MR, Thackeray R, Pass RF. Awareness of and behaviors related to child-to-mother transmission of cytomegalovirus. Prev Med 2012 May;54(5):351-357 [FREE Full text] [doi: 10.1016/j.ypmed.2012.03.009] [Medline: 22465669]
28. Congressional record - Senate S4106. United States Senate. 2011. URL: https://www.congress.gov/112/crec/2011/06/23/CREC-2011-06-23-pt1-PgS4106-3.pdf [accessed 2025-12-23]
29. eHealth: health topics. World Health Organization Regional Office for the Eastern Mediterranean. URL: https://www.emro.who.int/health-topics/ehealth/ [accessed 2024-12-16]
30. Social media in America: key stats in 2025. Sprinklr. 2025. URL: https://www.sprinklr.com/blog/social-media-in-america/ [accessed 2025-11-16]
31. Schumacher S, Sparks G, Montalvo III J, Kirzinger A, Hamel L. KFF health information and trust tracking poll: health information and advice on social media. KFF. 2025 Aug 7. URL: https://www.kff.org/public-opinion/kff-health-information-and-trust-tracking-poll-health-information-and-advice-on-social-media/ [accessed 2025-11-16]
32. Gottfried J. Americans’ social media use. Pew Research Center. 2024 Jan 31. URL: https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/ [accessed 2025-11-16]
33. June is national CMV awareness month - week 1 “CMV is common”. National CMV Foundation. 2016 Jun. URL: https://www.nationalcmv.org/cmv-research/blog/june-2016/national-cmv-awareness-month [accessed 2025-11-16]
34. About cytomegalovirus: cytomegalovirus (CMV) and congenital CMV infection. Centers for Disease Control and Prevention. 2025 Jan 17. URL: https://www.cdc.gov/cytomegalovirus/about/index.html [accessed 2025-11-16]
35. Singh M, Jakhar AK, Pandey S. Sentiment analysis on the impact of coronavirus in social life using the BERT model. Soc Netw Anal Min 2021;11(1):33 [FREE Full text] [doi: 10.1007/s13278-021-00737-z] [Medline: 33758630]
36. Mirugwe A, Ashaba C, Namale A, Akello E, Bichetero E, Kansiime E, et al. Sentiment analysis of social media data on Ebola outbreak using deep learning classifiers. Life (Basel) 2024 May 30;14(6):708 [FREE Full text] [doi: 10.3390/life14060708] [Medline: 38929691]

37. Ahmad M, Batyrshin I, Sidorov G. Sentiment analysis using a large language model–based approach to detect opioids mixed with other substances via social media: method development and validation. JMIR Infodemiology 2025 Jun 19;5(1):e70525 [FREE Full text] [doi: 10.2196/70525] [Medline: 40536906]
38. Deiner MS, Deiner NA, Hristidis V, McLeod SD, Doan T, Lietman TM, et al. Use of large language models to assess the likelihood of epidemics from the content of tweets: infodemiology study. J Med Internet Res 2024 Mar 01;26:e49139 [FREE Full text] [doi: 10.2196/49139] [Medline: 38427404]
39. Tastad KJ, Schleiss MR, Lammert SM, Basta NE. Awareness of congenital cytomegalovirus and acceptance of maternal and newborn screening. PLoS One 2019 Aug 26;14(8):e0221725 [FREE Full text] [doi: 10.1371/journal.pone.0221725] [Medline: 31449545]
40. Binda S, Pellegrinelli L, Terraneo M, Caserini A, Primache V, Bubba L, et al. What people know about congenital CMV: an analysis of a large heterogeneous population through a web-based survey. BMC Infect Dis 2016 Sep 26;16(1):513 [FREE Full text] [doi: 10.1186/s12879-016-1861-z] [Medline: 27671033]
41. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv Preprint posted online 2019 [FREE Full text] [doi: 10.48550/arXiv.1810.04805]
42. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174 [FREE Full text] [Medline: 843571]
43. Code of Federal Regulations: § 46.104 Exempt research. U.S. Department of Health and Human Services. 2026 Jan 16. URL: https://www.ecfr.gov/current/title-45/part-46/section-46.104#p-46.104(d)(4)(i)

Abbreviations

AI: artificial intelligence
API: application programming interface
BERT: Bidirectional Encoder Representations from Transformers
cCMV: congenital cytomegalovirus.
CMV: cytomegalovirus
LLM: large language model
NCMVF: National Cytomegalovirus Foundation

Edited by T Mackey; submitted 19.Jul.2025; peer-reviewed by M Marian, A Rasool; comments to author 30.Sep.2025; revised version received 10.Dec.2025; accepted 17.Dec.2025; published 22.Jan.2026.

Please cite as:
Rosebrock TR, Yang Z, D'Arco L, Pathak T, Vislay-Wade R, Fowler K, Diaz-Decaro J, Kunzweiler C
Using Artificial Intelligence Methods to Evaluate the Effect of the National Cytomegalovirus Awareness Month on the Content and Sentiment of Social Media Posts: Infodemiology Study
JMIR Infodemiology 2026;6:e80922
URL: https://infodemiology.jmir.org/2026/1/e80922
doi:10.2196/80922
PMID:

Quality, Reliability, and Dissemination of In Vitro Fertilization–Related Videos on Chinese Social Media: Cross-Sectional Analysis of 300 Short Videos

Dapeng Chu[1*], PhD; Xueyan Bai[1*], MD; Feng Guo[2*], BA

[1]Beijing Chao-Yang Hospital, Capital Medical University, Main Campus, Beijing Chao-Yang Hospital, 8 Gongren Tiyuchang Nanlu, Chaoyang District, Beijing, 100020, Beijing, China

[2]APUCH Innovation, Beijing, China

[*]all authors contributed equally

Corresponding Author:
Dapeng Chu, PhD
Beijing Chao-Yang Hospital, Capital Medical University, Main Campus, Beijing Chao-Yang Hospital, 8 Gongren Tiyuchang Nanlu, Chaoyang District, Beijing, 100020, Beijing, China

Abstract

Background: Patients increasingly rely on short-video platforms for information regarding in vitro fertilization (IVF), yet the relationship between the scientific quality of this content and its algorithmic dissemination remains unclear.

Objective: This study aimed to assess the quality, reliability, and key drivers of dissemination of IVF-related short videos on major Chinese social media platforms.

Methods: A cross-sectional content analysis was conducted on 300 popular IVF-related videos (the top 100 results from each platform) retrieved from Douyin, Bilibili, and Xiaohongshu between January 10 and 15, 2025. Video quality and reliability were evaluated using the Global Quality Score and a modified DISCERN instrument. Predictors of video dissemination were identified using an Extreme Gradient Boosting machine learning model, with the number of "likes" serving as the primary outcome variable.

Results: Content produced by medical professionals demonstrated significantly higher quality and reliability (median mDISCERN 11.0, IQR 9.0-15.0) compared to non-medical sources (median mDISCERN 8.0, IQR 5.0-13.0; $P< .001$). However, the Extreme Gradient Boosting analysis identified the uploader's follower count as the most powerful predictor of video "likes." In contrast, quality metrics (Global Quality Score and modified DISCERN scores) had a negligible impact on dissemination.

Conclusions: In the current Chinese social media landscape, the dissemination of IVF-related videos is strongly associated with creator influence rather than scientific merit. This disconnect between engagement and quality poses a potential risk of misinformation, highlighting the need for medical professionals to adopt platform-native communication strategies to ensure that high-quality information reaches patients.

(JMIR Infodemiology 2026;6:e83900) doi:10.2196/83900

KEYWORDS

in vitro fertilization; social media; health communication; content quality; misinformation

Introduction

In vitro fertilization (IVF) offers profound hope to individuals and couples facing infertility, yet the journey is fraught with challenges. The complexity of the procedures, significant financial costs, and uncertain outcomes impose a substantial physiological, psychological, and economic burden on patients [1]. In navigating this demanding process, access to accurate, comprehensive, and understandable medical information is critical for informed decision-making, managing treatment expectations, and mitigating psychological distress [2-5]. Historically, this information was primarily disseminated by health care institutions. However, the digital era has precipitated a paradigm shift, with patients increasingly turning to the internet for more accessible and diverse sources of support [6-8]. While social media's impact on health behaviors is a global phenomenon, China's digital ecosystem offers a unique context for study. With the world's largest internet user base and high demand for assisted reproductive technology, platforms such as Douyin (the Chinese counterpart to TikTok) provide a critical "natural laboratory" to understand algorithmic health communication patterns that are increasingly relevant worldwide.

In recent years, short-video platforms such as Douyin, Bilibili, and Xiaohongshu have emerged as dominant arenas for health information dissemination, distinguished by their algorithm-driven, highly engaging, and rapidly propagating

nature [9,10]. While these platforms present unprecedented opportunities for medical education, they also introduce formidable challenges [11-13]. Unlike traditional medical websites, content generation is often spontaneous and lacks rigorous professional oversight, creating a "perfect storm of information" where quality is highly variable [14,15]. Furthermore, their personalized recommendation algorithms, while enhancing user experience, risk creating "information cocoons" that can amplify biased or inaccurate content [16,17], posing a potential hazard to patients seeking IVF treatment, especially concerning misinformation on reproductive health [18-20].

Despite growing analyses of social media health content, the IVF domain on Chinese short-video platforms remains understudied. Moreover, prior work has largely assessed quality in isolation, leaving unclear whether intrinsic quality or extrinsic platform factors (eg, creator influence) primarily drive dissemination. The few studies that have explored dissemination dynamics have relied on conventional linear models, which are ill-equipped to capture the complex, nonlinear factors driving content virality in sophisticated social networks. This leaves our understanding of the contemporary health information ecosystem fundamentally incomplete.

To address this gap, we conducted a cross-sectional analysis of the top-ranked IVF videos on Douyin, Bilibili, and Xiaohongshu. Our methodological approach consisted of three phases: (1) content analysis to classify uploader identity and topics, (2) assessment of information quality using the Global Quality Score (GQS) and the modified DISCERN (mDISCERN) instrument, and (3) machine learning analysis using Extreme Gradient Boosting (XGBoost) and Shapley Additive Explanations (SHAP) values to isolate independent predictors of video dissemination among metadata variables. We hypothesized that uploader influence (follower count), rather than content quality, would be the dominant predictor of engagement. The findings are intended to provide an evidence-based foundation for enhancing the effective communication of high-quality medical information and to offer actionable guidance for platforms, content creators, and public health authorities.

Methods

Study Design and Video Retrieval

A cross-sectional study was designed to evaluate the quality, reliability, and dissemination of IVF-related videos across three popular Chinese social media platforms: Xiaohongshu, Bilibili, and Douyin [21]. These platforms were selected based on their market dominance and distinct demographic profiles. Douyin (Chinese version of TikTok, (ByteDance, Beijing, China), with 766 million daily active users (DAUs) as of 2024 [22], represents China's short-video mainstream platform, with videos typically ranging from 15 to 60 seconds in length [23], Bilibili (Bilibili Inc., Shanghai, China), with an average of 104 million DAUs in 2024 [24], specializes in medium-to-long form video content (typically 3 - 30 min), with medium and long videos accounting for 70% of platform views [25]. The platform's user base is predominantly young, with nearly 70% of China's Generation Z population and an average user age of 25 years [24]. Xiaohongshu (RedNote, Xiaohongshu, Shanghai, China), with 143 million global DAUs by the end of 2024 [26], serves as a lifestyle-focused platform with a predominantly female user base (70% female) and functions as a primary search engine for lifestyle and health-related decisions among Chinese women [27].

Using the Chinese keyword "试管婴儿" (IVF), relevant videos were retrieved from each platform between January 10 and 15, 2025. To mitigate the influence of personalized recommendations, searches were conducted using newly created accounts with no prior viewing history. No filters or sorting mechanisms were applied, thereby simulating a typical user experience.

An initial systematic search on the 3 platforms identified 531 potentially relevant videos. These records were then screened for eligibility based on predefined inclusion and exclusion criteria. A total of 231 videos were excluded for the following reasons: duplicate content (n=127, 55.0%), non–Chinese-language content (n=14, 6.1%), being purely promotional without educational value (n=55, 23.8%), or having content irrelevant to the topic of IVF (n=35, 15.2%). This screening process yielded a final sample of 300 (56.5%) unique videos for analysis, comprising the top 100 eligible videos from each platform.

We used a quota sampling strategy based on platform search rankings. For each platform, videos were retrieved and screened sequentially, starting from the top-ranked search result. The screening process continued down the ranked list until a quota of 100 eligible videos meeting all inclusion and exclusion criteria was reached for each platform. This strategy ensures that the sample reflects the content most visible to users, as search rankings prioritize high-engagement content. These rankings are algorithmically driven and prioritize a synthesis of user engagement metrics (eg, likes, comments, and shares), topical relevance, and content freshness, thereby simulating the ecological search experience of a typical user. Selection was subject to the following criteria: (1) the video was in the Chinese language, (2) the video focused on IVF-related medical content, and (3) the video was publicly accessible. Videos were excluded if they contained (1) duplicate content, (2) pure advertisements without educational value, (3) videos unrelated to IVF, and (4) non–Chinese-language videos. For each video, basic information was documented, including title, upload date, duration, uploader identity, and engagement metrics (eg, likes, comments, shares, and saves). Regarding cross-platform posting, videos uploaded by the same creator to multiple platforms were treated as distinct analytical units. This approach was chosen because engagement metrics (eg, likes and comments) are platform specific and reflect the unique algorithmic distribution and audience reaction within that specific ecosystem. However, intraplatform duplicates (the same video uploaded twice to the same platform) were excluded.

The diagram illustrates the selection process for the study. Initially, a keyword search for "试管婴儿" (IVF) was conducted on 3 platforms, namely, Douyin, Bilibili, and Xiaohongshu, identifying 231, 500, and 390 videos, respectively. Following

the platforms' comprehensive ranking algorithms, the top-ranked videos were selected for screening (Douyin, n=116; Bilibili, n=120; and Xiaohongshu, n=122). During the screening phase, videos were excluded for being duplicates or thematically irrelevant (Douyin, n=16; Bilibili, n=20; and Xiaohongshu, n=22). Finally, 100 eligible videos from each platform were included, resulting in a total of 300 videos for the final analysis.

Video Classification

Videos were categorized by uploader and content type by 2 independent researchers, with disagreements resolved by a third reviewer, following established content analysis methodologies [28,29].

All included videos were systematically classified according to a predefined coding scheme focusing on two primary dimensions: uploader identity and content theme. The uploader of each video was categorized into one of five groups: medical professionals, which included verified IVF doctors, reproductive medicine institutions, or health care providers; health science communicators or key opinion leaders, defined as individuals known for disseminating health knowledge without formal medical credentials; patients and sharers, consisting of individuals sharing personal IVF treatment experiences; marketing promoters, identified as commercial entities promoting fertility services; and news and general content creators, such as media outlets providing general information.

Concurrently, the primary subject matter of each video was assigned to one of five content categories: medical knowledge, comprising scientific explanations, clinical guidelines, or technical information; fertility and lifestyle optimization, focusing on content related to lifestyle practices intended to enhance fertility; patient experience sharing, which covered personal narratives detailing individual IVF journeys; policy and ethical topics, which included discussions on regulations or social implications; and misleading or marketing content, which included promotional material or videos with verifiably inaccurate information.

Quality and Reliability Assessment

Video quality and reliability were assessed using the GQS [30,31] and the mDISCERN instrument [32,33].

The GQS is a 5-point scale evaluating overall quality, flow, and integrity of information (1=poor and 5=excellent) and has been validated in numerous studies of web-based health information [30,31]. The mDISCERN instrument was adapted from the original 16-item DISCERN tool [32,33] to specifically evaluate short-form digital health content. To accommodate the brevity of social media videos, the instrument was condensed into five core dimensions:

1. Reliability of information: assesses the evidentiary basis and accuracy of medical claims
2. Clarity of aims: evaluates whether the video's purpose and structure are clearly communicated
3. Relevance of sources: measures the transparency and authority of cited evidence (eg, clinical guidelines vs unverifiable anecdotes)
4. Balance and impartiality: assesses the extent of commercial bias or one-sided promotion
5. Presentation of uncertainty: evaluates the disclosure of risks, side effects, and biological variability

Items 9 to 15 of the original DISCERN tool, which focus on detailed treatment choices and shared decision-making, were excluded as they are rarely applicable to brief, nonconsultative video clips. Furthermore, the scoring system was modified from the original 1- to 5-point scale to a 0- to 5-point Likert scale. This modification allowed for a score of "0" to explicitly categorize content that was completely devoid of sources, reliability, or clear aims, which is a common characteristic of low-quality user-generated content.

Two experienced reproductive medicine specialists independently scored all videos. Initial disagreements were resolved through discussion to reach a consensus. If a consensus could not be reached, the score from the third senior specialist was used as the final arbitration score. Interrater reliability was assessed and found to be high (Cohen κ coefficient >0.80), indicating a strong degree of agreement [34,35].

Dissemination Analysis With XGBoost

An XGBoost regression model was used to identify factors influencing video dissemination, with the number of "likes" as the primary outcome variable [36,37]. Input features included platform, uploader category, content category, GQS score, mDISCERN score, video length, days since publishing, and the presence of background music or subtitles. To investigate the primary drivers of raw engagement, the model was first trained on the untransformed "likes" count.

Furthermore, to account for the highly skewed distribution of the "likes" variable and to build a more stable model for confirming feature contributions, a secondary analysis was performed using a logarithmic transformation (log1p) on the target variable before model training. This standard preprocessing technique helps mitigate the influence of outlier videos with extremely high engagement.

Hyperparameter tuning was performed using grid search with 3-fold cross-validation to optimize model performance [38,39]. The importance of features was assessed using SHAP values to provide transparent and intuitive insights into how each factor influences dissemination [40,41].

Statistical Analysis

All statistical analyses were conducted using R (version 4.2.3; R Foundation for Statistical Computing). Owing to the nonnormal distribution, nonparametric tests were used, including the Kruskal-Wallis test for multigroup comparisons and the Dunn test for post-hoc analysis. The Spearman correlation coefficient was used to assess relationships between variables. Statistical significance was defined as $P<.05$.

Ethical Considerations

This study analyzed publicly available, user-generated IVF-related videos and their metadata from Chinese social media platforms. The research involved no interaction or intervention with individuals and did not collect, store, or process identifiable

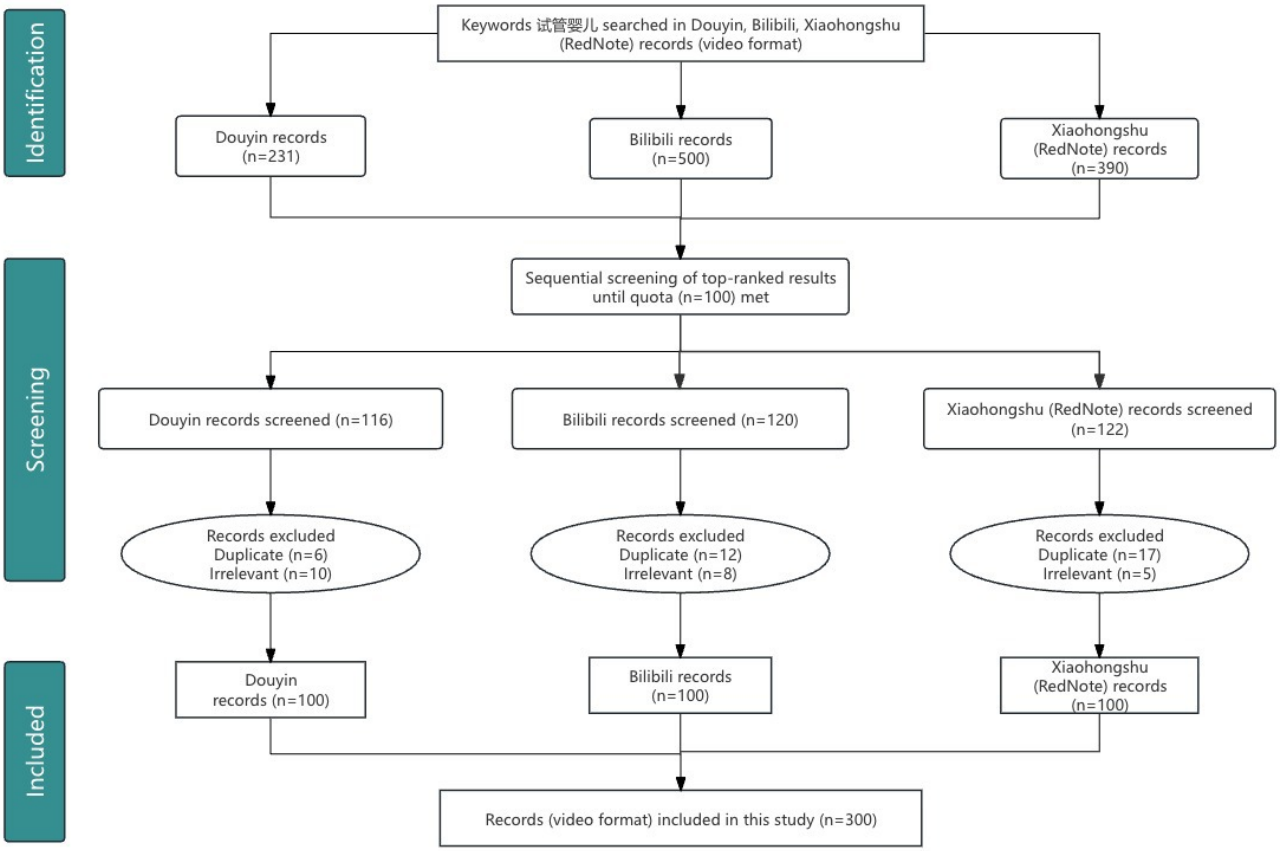
personal data. In accordance with institutional and national guidelines, the use of publicly accessible, aggregate data does not constitute human subjects research; therefore, ethics approval and consent to participate were not required.

Results

Overview

The complete selection process is detailed in the video selection flow diagram (Figure 1). The study's findings reveal a fundamental paradox in the digital health ecosystem: while the quality of IVF-related information is critically dependent on the professional identity of its creator, its dissemination is overwhelmingly governed by the creator's platform influence, not the quality of the content itself.

Figure 1. Flow diagram of the video identification, screening, and inclusion process.



Platform Ecosystems Exhibit Profound Heterogeneity

Analysis of 300 videos revealed significant heterogeneity in content strategy, user base, and engagement dynamics across platforms (P<.001; Table 1). On Douyin and Xiaohongshu, medical professionals were the dominant uploaders (87/100, 87%, and 89/100, 89% respectively), and content was primarily "medical knowledge" (71/100, 71% on both). In contrast, Bilibili featured a more diverse creator base, including patient sharers (28/100, 28%) and health science communicators (22/100, 22%), with "patient experience sharing" (31/100, 31%) being more prevalent (Table 2).

Table . Baseline of in vitro fertilization–relevant videos.

Characteristics	Douyin (n=100)	Bilibili (n=100)	Xiaohongshu (n=100)	*P* value
Likes, median (IQR)	539 (81 - 2986.5)	138 (16 - 844.5)	277.5 (66 - 925.25)	<.001
Saves, median (IQR)	145 (11-732)	78 (12-415)	134 (40 - 525.75)	.31
Comments, median (IQR)	30 (6.75 - 327.25)	17 (3 - 149.5)	51 (7.75 - 126.5)	.16
Shares, median (IQR)	161 (9.5 - 785)	56 (5-310)	96.5 (26.5 - 437.75)	.10
Days since uploading, median (IQR)	116 (100.75 - 155)	788.5 (293.25 - 1300.75)	218.5 (125 - 394.25)	<.001
Length, median (IQR)	41.5 (25.75 - 59)	325 (169-692)	53 (37.75 - 94)	<.001
Followers, median (IQR)	36000 (7469.25 - 273,500)	4679.5 (322.25 - 35,500)	8769.5 (3010.75 - 33,000)	<.001
Total video count, median (IQR)	364 (167-560)	201 (66.5 - 488.5)	309.5 (191.25 - 758.25)	.004
Type of video, n				<.001
Medical knowledge	71	43	71	
Fertility & lifestyle optimization	0	10	3	
Patient experience sharing	19	31	18	
Policy & ethical topics	8	12	3	
Misleading or marketing content	2	4	5	
Type of uploader, n				<.001
Medical professionals	87	24	89	
Health science communicators/medical key opinion leaders (KOLs)	1	22	0	
Patients and fertility journey sharers	7	28	5	
Marketing promoters	1	11	6	
News and general interest content creators	4	15	0	
BGM[a], n				<.001
Without BGM	38	46	19	
With BGM	62	54	81	
Subtitle, n				.02
Without subtitle	0	6	1	
With subtitle	99	94	99	
GQS[b] score, median (IQR)	2 (2-3)	2 (2-3)	3 (2-3)	.04
DISCERN score, median (IQR)	10 (7-14)	10 (6 - 14.25)	12 (9-15)	.02

[a]BGM: background music.

[b]GQS: Global Quality Score.

Table . Characteristics of uploaders and video content across the 3 platforms (N=300).

Category	Douyin (n=100), n (%)	Bilibili (n=100), n (%)	Xiaohongshu (RedNote) (n=100), n (%)	*P* value
Uploader profile				<.001
Medical professionals	87 (87)	24 (24)	89 (89)	
Patient voices	1 (1)	28 (28)	5 (5)	
Health communicators or KOLs[a]	7 (7)	22 (22)	6 (6)	
Marketing and promoters	4 (4)	11 (11)	0 (0)	
News and general content	1 (1)	15 (15)	0 (0)	
Video content type				<.001
Medical knowledge	71 (71)	43 (43)	71 (71)	
Patient experience	19 (19)	31 (31)	18 (18)	
Fertility and lifestyle	8 (8)	12 (12)	3 (3)	
Policy and ethics	0 (0)	10 (10)	3 (3)	
Misleading and marketing	2 (2)	4 (4)	5 (5)	

[a]KOL: key opinion leader.

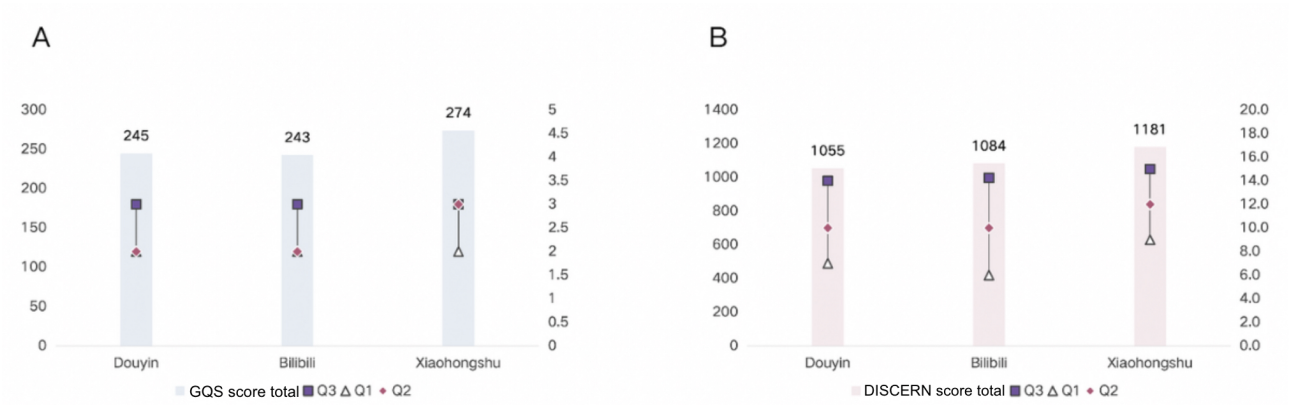
Video attributes also differed significantly. Bilibili hosted older (median 788.5, IQR 293.3-1300.8 d) and longer videos (median 325, IQR 169-692 s), whereas Douyin featured more recent, shorter-form content (median 41.5, IQR 25.8-59 s; *P*<.001 for both). Douyin creators had the largest median follower counts (n=36,000) and generated the highest median "likes" (n=539), significantly outperforming the other platforms. These baseline differences in creator influence and content strategy precede the analysis of dissemination drivers.

Quality and Reliability Assessment

Overall, video quality and reliability were moderate. Interrater agreement for the scoring was high (weighted κ_{GQS}=0.82, 95% CI 0.76 - 0.87; weighted $\kappa_{mDISCERN}$=0.79, 95% CI 0.72 - 0.85; International Code Council$_{GQS}$=0.90, 95% CI 0.86 - 0.93; International Code Council$_{mDISCERN}$=0.88, 95% CI 0.84 - 0.91).

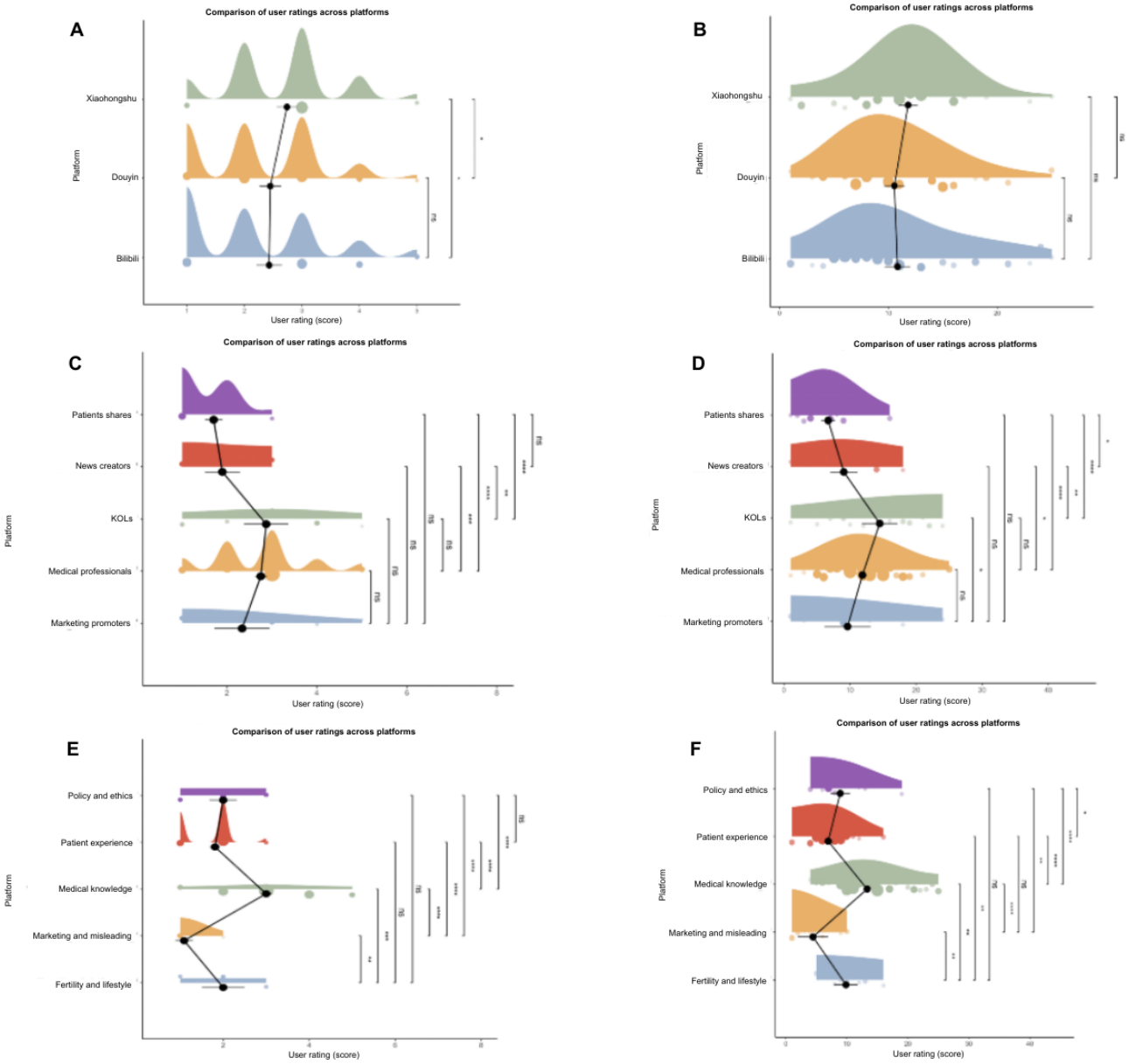
Platform-level analysis showed that Xiaohongshu videos achieved a statistically significant higher quality, with a median GQS of 3.0 (IQR 2.0-3.0; *P*=.04) and a median mDISCERN score of 12.0 (IQR 9.0-15.0; *P*=.02) compared to Douyin and Bilibili (Figure 2).

Figure 2. Comparison of Global Quality Score (GQS) and DISCERN score distributions across 3 platforms. (A) Distribution of GQS scores across the 3 platforms; (B) distribution of DISCERN scores across the 3 platforms.



Subgroup analysis revealed a strong association between uploader identity and content quality (*P*<.001). Our data highlights a distinct "competence hierarchy": videos from medical professionals consistently achieved the highest reliability scores (median mDISCERN 11.0, IQR 9.0-15.0), reflecting adherence to clinical guidelines. Conversely, patient sharers and marketing promoters scored significantly lower. While patient sharers provide emotional value, their content often lacked medical accuracy (median GQS 2.0, IQR 1.0-2.0), suggesting that the "lived experience" often comes at the expense of clinical precision. Videos categorized as "medical knowledge" were rated significantly higher than all other content themes (*P*<.001; Figure 3).

Figure 3. Comparison of video quality and reliability scores across different subgroups. (A) Distribution of Global Quality Score (GQS) scores across the 3 platforms; (B) distribution of DISCERN scores across the 3 platforms; (C) distribution of GQS scores across the 5 uploader types; (D) distribution of DISCERN scores across the 5 uploader types; (E) distribution of GQS scores across the 5 content types; and (F) distribution of DISCERN scores across the 5 content types. Each subplot displays the data distribution using a violin plot and a box plot. Asterisks indicate the level of statistical significance from pairwise statistical tests. *P<.05, **P<.01, ***P<.001. KOL: key opinion leader.

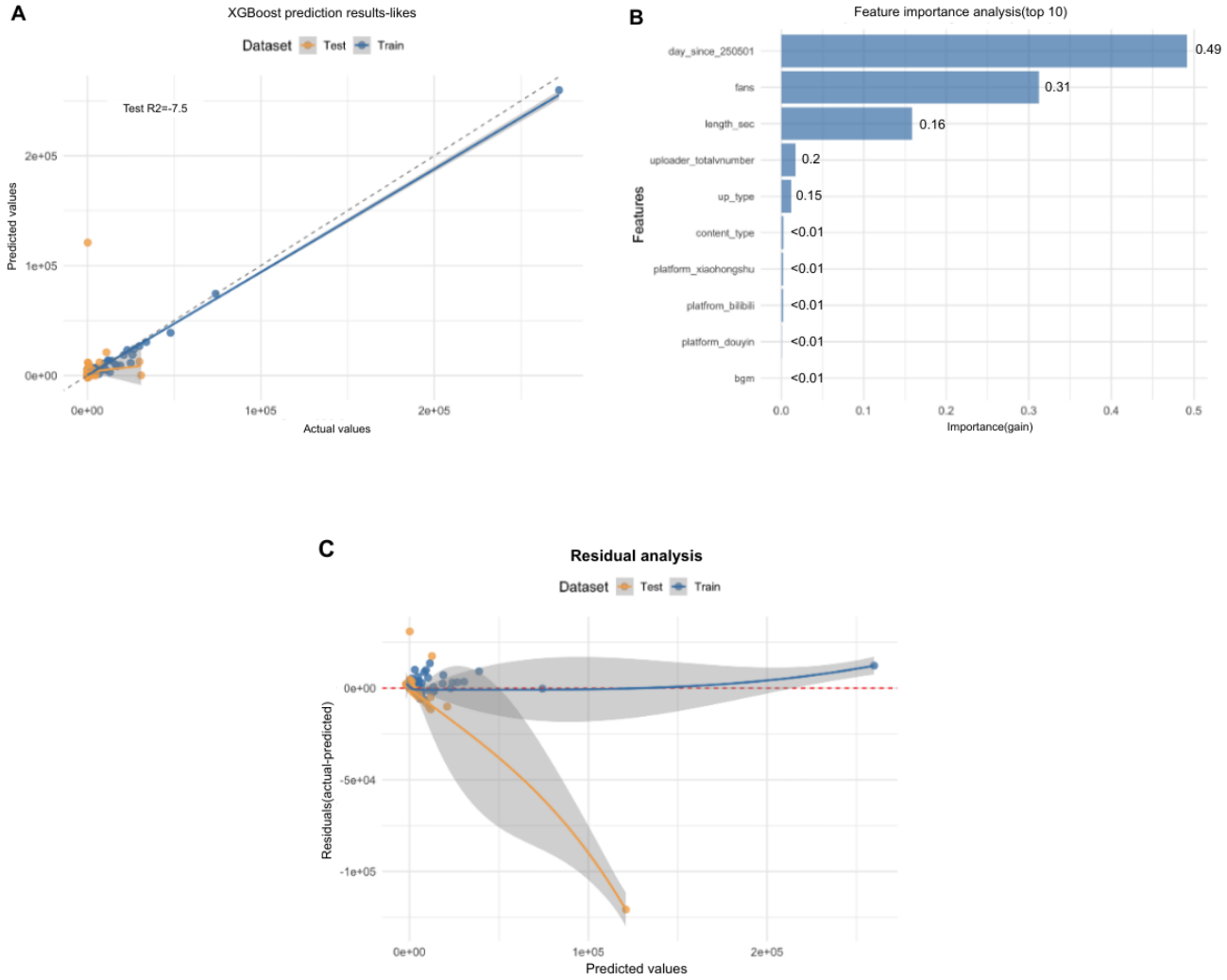


Predictors of Video Dissemination

To resolve the disconnect between content creation and consumption, an XGBoost machine learning model was developed to identify the primary drivers of video dissemination ("likes"). Initial bivariate correlation analysis found no significant relationship between a video's like count and its GQS or mDISCERN scores, providing a preliminary suggestion that quality was not a key factor for engagement.

The initial XGBoost model, trained directly on the untransformed "likes" count, yielded a negative R^2 of −7.5 (Figure 4A). This result confirms that raw social media engagement follows a nonlinear, heavy-tailed distribution that cannot be modeled by standard additive regression. Consequently, feature importance rankings from this initial model were disregarded to avoid spurious conclusions (Figure 4). We therefore relied exclusively on the secondary log1p-transformed model for identifying predictors.

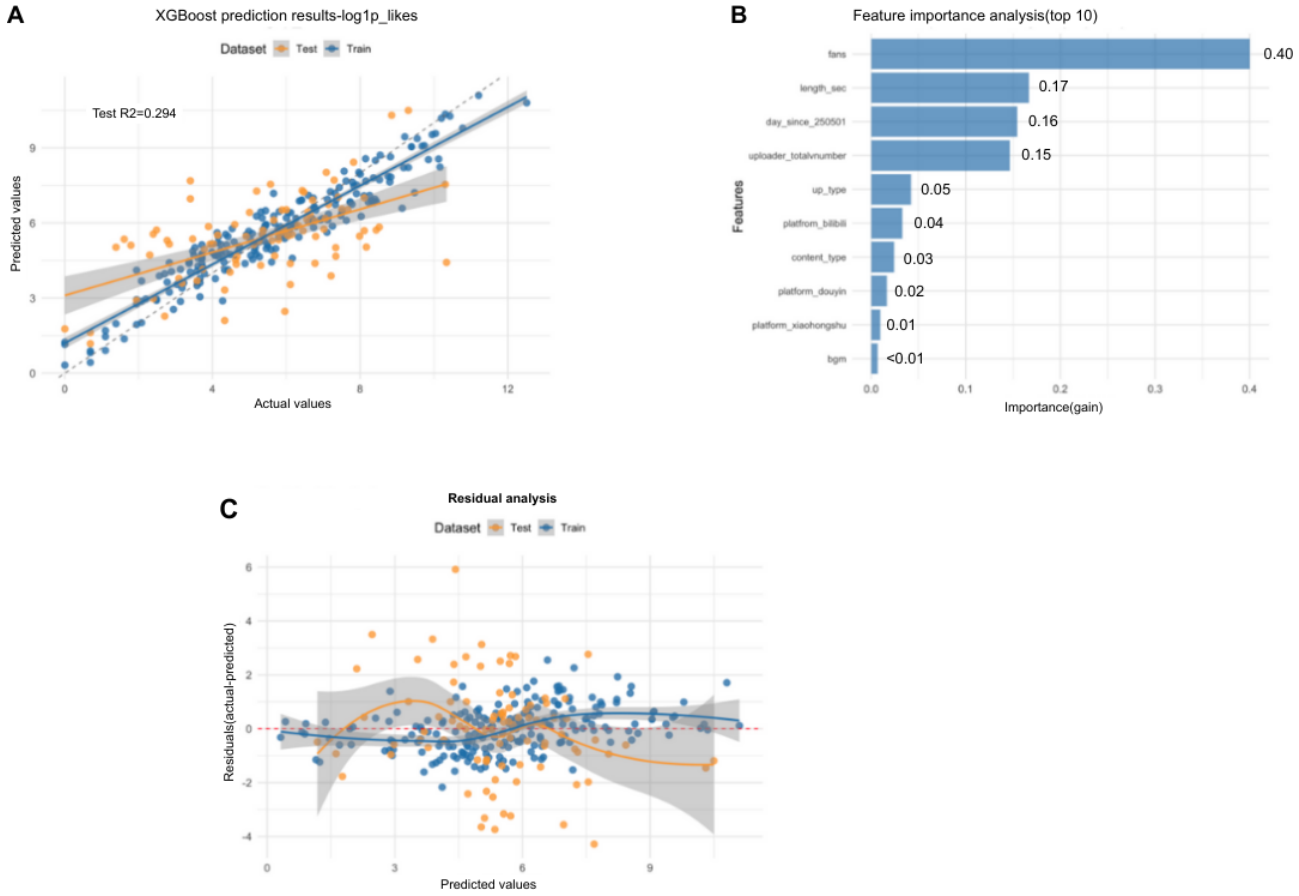
Figure 4. Results of the XGBoost model interpreting the factors influencing video likes: (A) the goodness of fit of the XGBoost model (R^2=−7.5); (B) the importance of each variable on the number of video likes; and (C) the Shapley Additive Explanation summary plot showing the impact of important variables on the model output. XBoost: Extreme Gradient Boosting.



A closer examination of the SHAP summary plot (Figure 4) clarifies how these top features influence predictions. For the "fans" feature, a clear positive trend is visible: high feature values (represented by red dots) are strongly associated with high positive SHAP values, confirming that a larger follower count directly contributes to a higher prediction of likes. In contrast, for metrics such as GQS and mDISCERN, the points are clustered vertically around the zero-line with no discernible color gradient, visually confirming their negligible impact on the model's output and reinforcing the quality-impact gap.

To ensure these findings were not an artifact of the target variable's extreme skewness and to robustly validate the feature hierarchy, we developed a second XGBoost model by applying a logarithmic transformation (log1p) to the "likes" count. This standard data preprocessing step resulted in a more stable model with a positive predictive fit, achieving an R^2 of 0.294 on the test set (Figure 5A). The corresponding residual plot confirmed a more balanced and desirable error distribution (Figure 5C).

Figure 5. Performance and interpretation of the XGBoost model for predicting video likes. (A) Prediction results of the XGBoost model, plotting actual versus predicted values for the log1p_likes target. The coefficient of determination for the test set is R^2=0.294. (B) Feature importance analysis showing the top 10 predictors ranked by their gain score. (C) Residual analysis plotting residuals against predicted values to assess model fit and error distribution. XGBoost: Extreme Gradient Boosting.



Crucially, despite the improved model performance, the feature importance analysis remained remarkably consistent (Figure 5B). The uploader's follower count was once again the most dominant predictor. In stark contrast, the validated metrics for content quality and reliability—GQS and mDISCERN scores—remained at the bottom of the feature importance hierarchy, exerting negligible influence. This dual-model approach provides robust, 2-fold evidence that in the current algorithmic landscape, a video's reach is driven primarily not by its scientific merit but by the preexisting social capital of its creator.

Discussion

The "Quality-Impact Gap" in Digital Health

This study provides the first systematic evaluation of IVF-related health information on major Chinese short-video platforms, revealing a significant and troubling paradox at the heart of the modern digital health landscape. Our analysis empirically demonstrates a divergence between content quality and engagement. While quality metrics (GQS and mDISCERN) track closely with medical expertise, dissemination metrics (likes) track with uploader influence. This suggests that high-quality medical information does not automatically generate high engagement. This "quality-impact gap" [42] implies that scientific accuracy is not the primary driver of algorithmic visibility. This phenomenon is not merely an algorithmic quirk; it reflects a fundamental tension between the clinical nature of information and the socioemotional needs of patients. Patients navigating the arduous IVF journey are not just passive consumers of data; they are actively seeking hope, validation, and a sense of community. Consequently, low-quality but emotionally resonant content—such as unverified "miracle baby" testimonials—may be perceived as more valuable than dry, technically accurate explanations, leading to higher engagement.

Clinically, this gap risks therapeutic misconception and spending on unproven add-ons among IVF patients. To mitigate harm, clinicians and fertility centers should coproduce platform-native content—short, narrative-driven videos that embed evidence (success rates and indications or contraindications), use on-screen references, and include myth-fact segments—and deploy them via verified accounts with regular posting cadence and call-to-action links to authoritative resources [11,43].

Our central finding—that an uploader's follower count is the most potent predictor of reach—must be interpreted with nuance. In the fast-paced digital environment, follower count acts as a powerful cognitive heuristic for trust. Lacking the time or expertise to critically appraise every video, users subconsciously substitute "popularity" for "credibility," operating under the assumption that a large following implies authority and trustworthiness [43]. This dynamic, where social capital eclipses scientific capital, aligns perfectly with findings from Western

platforms such as YouTube [44,45] and extends recent analyses of other medical topics on Chinese platforms [46,47]. Our research confirms that this algorithmic prioritization of engagement over evidence is a universal feature of contemporary social media architecture [48,49], creating a global challenge for evidence-based health communication.

Platform Ecosystems: Expert-Led Versus Community-Driven

Furthermore, our 3-platform comparison revealed distinct "platform personalities" that shape this information flow. Douyin and Xiaohongshu function primarily as expert-led, knowledge-dissemination channels, yet they are still subject to the influencer dynamic. In contrast, Bilibili operates as a community-driven, experience-sharing hub (with 28% patient sharers vs ≤7% on other platforms), hosting longer narratives that fulfill patients' documented need for peer support [50]. While valuable for emotional well-being, this "experiential" content was found to be of significantly lower quality, posing a risk of normalizing anecdotal advice over clinical guidelines.

The implications of this ecosystem for patient care and public health are profound. For a vulnerable population already facing immense emotional and financial stress, the stakes are exceptionally high. Exposure to misinformation or low-quality content can foster therapeutic misconceptions, leading patients to pursue unproven and costly adjunct therapies. It can create unrealistic expectations about success rates, leading to deeper psychological distress when treatments fail. Moreover, the prevalence of marketing content masquerading as educational material exposes patients to potential financial exploitation [51]. The "attention economy" of social media is thus not a neutral marketplace of ideas; for IVF patients, it is a high-risk environment where the most visible information is often the least reliable.

Clinical Implications and Future Directions

Therefore, a paradigm shift is imperative for medical professionals and health care organizations. A passive approach of simply producing high-quality content and expecting it to be discovered is destined to fail. A proactive, 2-pronged strategy is required. First, proactive content creation demands that clinicians become platform-native communicators [11,52,53]. This means moving beyond static informational videos and embracing storytelling, patient-centered narratives, and visually compelling formats developed in collaboration with communication experts, without compromising scientific integrity [54]. Second, reactive engagement is equally crucial. Medical professionals and institutions should consider themselves "digital first responders," actively identifying and correcting high-reach misinformation through comments, response videos, or collaborations with platforms—a practice shown to be effective in other health contexts [43].

Looking forward, a clear agenda for future research emerges from this work. While our quantitative model identified what drives dissemination, qualitative studies are now needed to understand why. In-depth interviews with IVF patients could illuminate the specific motivations and cognitive processes behind their engagement with different types of content. Furthermore, interventional research is urgently needed to design and test the efficacy of novel communication strategies. Randomized controlled trials could compare the reach and impact of standard informational videos against narrative-based, emotionally resonant, yet scientifically accurate content. Finally, longitudinal studies are required to track the real-world impact of social media exposure on patient decision-making, treatment adherence, and clinical outcomes over time.

This study has several notable strengths. It is the first to systematically analyze IVF content across China's 3 dominant short-video platforms. By using a robust content analysis methodology underpinned by validated instruments—GQS and the mDISCERN tool with high interrater reliability (κ>0.80)—we provided a rigorous assessment of content quality. Methodologically, our use of a dual XGBoost modeling strategy provides a particularly robust analysis. We first demonstrated the model's inability to predict absolute "likes" (R^2=−7.5), empirically confirming the highly stochastic nature of social media virality [55]. We then solidified our conclusions by using a second, logarithmically transformed model, which, despite a better predictive fit (R^2=0.294), reproduced the exact same feature importance hierarchy. This confirmatory step ensures that our central finding is robust and not an artifact of data skewness. Third, while our optimized XGBoost model achieved a robust R^2 of 0.294, approximately 70% of the variance in dissemination remains unexplained. This suggests that engagement is influenced by unmeasured "soft" factors extrinsic to medical quality or uploader status, such as thumbnail aesthetics, opening "hooks" (the first 3 s of video), emotional delivery, and platform-specific trending audio. Future studies should use computer vision and sentiment analysis to quantify these variables.

Limitations

However, the study's limitations must be acknowledged. First, our sampling strategy restricted analysis to the top-ranked videos. While this design validly represents the "information diet" of a typical user who rarely scrolls beyond the first page, it introduces selection bias. Our findings characterize the most visible content ecosystem rather than the entire universe of IVF-related videos. Additionally, its cross-sectional design precludes any causal inference. The findings are specific to the Chinese social media context and may not be generalizable. Finally, our analysis used "likes" as the primary proxy for dissemination. Future research could conduct a more granular analysis exploring the distinct drivers of other interaction types, such as comments or shares, which may reflect different dimensions of user engagement [56].

Conclusions

In conclusion, our research paints a stark picture of the IVF information landscape on Chinese social media, where the mechanisms of dissemination are dangerously decoupled from the principles of evidence-based medicine. Our dual-model analysis robustly demonstrates that the digital health ecosystem does not inherently reward quality; it rewards influence. To bridge the gap between what is popular and what is reliable, the medical community must not only produce trustworthy

information but also master the art and science of platform-native communication to ensure that their expertise can successfully navigate the algorithmic currents and reach the patients who need it most.

Funding

This work was supported by the National Natural Science Foundation of China (grant 82301803). The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Wyns C, Bergh C, Calhaz-Jorge C, et al. Assisted reproductive technology in Europe, 2016: results generated from European registries by ESHRE. Hum Reprod Open 2020(3):hoaa032. [doi: 10.1093/hropen/hoaa032]
2. Leone D, Menichetti J, Barusi L, et al. Breaking bad news in assisted reproductive technology: a proposal for guidelines. Reprod Health 2017 Jul 20;14(1):87. [doi: 10.1186/s12978-017-0350-1] [Medline: 28728610]
3. Mayette E, Scalise A, Li A, McGeorge N, James K, Mahalingaiah S. Assisted reproductive technology (ART) patient information-seeking behavior: a qualitative study. BMC Womens Health 2024 Jun 15;24(1):346. [doi: 10.1186/s12905-024-03183-z] [Medline: 38877503]
4. Lemoine ME, O'Connell SBL, Grunberg PH, Gagné K, Ells C, Zelkowitz P. Information needs of people seeking fertility services in Canada: a mixed methods analysis. Health Psychol Behav Med 2021 Feb 11;9(1):104-127. [doi: 10.1080/21642850.2021.1879650] [Medline: 34104552]
5. Karami NA, Latifi M, Berahmand N, Eini F, Al-Suqri MN. The impact of individual factors on health information-seeking behavior of infertile couples undergoing assisted reproductive technologies: Longo model. Adv Biomed Res 2023;12(1):68. [doi: 10.4103/abr.abr_181_22] [Medline: 37200740]
6. Slauson-Blevins KS, McQuillan J, Greil AL. Online and in-person health-seeking for infertility. Soc Sci Med 2013 Dec;99:110-115. [doi: 10.1016/j.socscimed.2013.10.019] [Medline: 24355477]
7. Jones CA, Mehta C, Zwingerman R, Liu KE. Fertility patients' use and perceptions of online fertility educational material. Fertil Res Pract 2020;6:11. [doi: 10.1186/s40738-020-00083-2] [Medline: 32695432]
8. Moran G, Muzellec L, Johnson D. Message content features and social media engagement: evidence from the media industry. J Prod Brand Manag 2019 Oct 16;29(5):533-545. [doi: 10.1108/JPBM-09-2018-2014]
9. Zeng F, Zhang W, Wang M, Zhang H, Zhu X, Hu H. Douyin and Bilibili as sources of information on lung cancer in China through assessment and analysis of the content and quality. Sci Rep 2024 Sep 4;14(1):20604. [doi: 10.1038/s41598-024-70640-y] [Medline: 39232044]
10. Gao H, Yin H, Peng L, Wang H. Effectiveness of social video platforms in promoting COVID-19 vaccination among youth: a content-specific analysis of COVID-19 vaccination topic videos on Bilibili. Risk Manag Healthc Policy 2022;15:1621-1639. [doi: 10.2147/RMHP.S374420] [Medline: 36071816]
11. Ventola CL. Social media and health care professionals: benefits, risks, and best practices. P T 2014 Jul;39(7):491-520. [Medline: 25083128]
12. Cain J. Social media in health care: the case for organizational policy and employee education. Am J Health Syst Pharm 2011 Jun 1;68(11):1036-1040. [doi: 10.2146/ajhp100589] [Medline: 21593233]
13. Chretien KC, Kind T. Social media and clinical care: ethical, professional, and social implications. Circulation 2013 Apr 2;127(13):1413-1421. [doi: 10.1161/CIRCULATIONAHA.112.128017] [Medline: 23547180]
14. Fernández-Luque L, Bau T. Health and social media: perfect storm of information. Healthc Inform Res 2015 Apr;21(2):67-73. [doi: 10.4258/hir.2015.21.2.67] [Medline: 25995958]
15. Laranjo L, Arguel A, Neves AL, et al. The influence of social networking sites on health behavior change: a systematic review and meta-analysis. J Am Med Inform Assoc 2015 Jan;22(1):243-256. [doi: 10.1136/amiajnl-2014-002841] [Medline: 25005606]
16. Sommariva S, Vamos C, Mantzarlis A, Đào LUL, Martinez Tyson D. Spreading the (fake) news: exploring health messages on social media and the implications for health professionals using a case study. Am J Health Educ 2018 Jul 4;49(4):246-255. [doi: 10.1080/19325037.2018.1473178]
17. Tasnim S, Hossain MM, Mazumder H. Impact of rumors and misinformation on COVID-19 in social media. J Prev Med Public Health 2020 May;53(3):171-174. [doi: 10.3961/jpmph.20.094] [Medline: 32498140]
18. Abbasi J. Widespread misinformation about infertility continues to create COVID-19 vaccine hesitancy. JAMA 2022 Mar 15;327(11):1013-1015. [doi: 10.1001/jama.2022.2404] [Medline: 35191947]

19. John JN, Gorman S, Scales D, Gorman J. Online misleading information about women's reproductive health: a narrative review. J Gen Intern Med 2025 Apr;40(5):1123-1131. [doi: 10.1007/s11606-024-09118-6] [Medline: 39511120]
20. Zaila KE, Osadchiy V, Shahinyan RH, Mills JN, Eleswarapu SV. Social media sensationalism in the male infertility space: a mixed methodology analysis. World J Mens Health 2020 Oct;38(4):591-598. [doi: 10.5534/wjmh.200009] [Medline: 32378368]
21. Olsen C, George DMM. Cross-sectional study design and data analysis. : College Entrance Examination Board; 2004 URL: http://www.yes-competition.org/media.collegeboard.com/digitalServices/pdf/yes/4297_MODULE_05.pdf [accessed 2025-12-24]
22. TikTok revenue and usage statistics (2025). Business of Apps. 2024. URL: https://www.businessofapps.com/data/tik-tok-statistics/ [accessed 2025-12-24]
23. Lu X, Lu Z. Fifteen seconds of fame: a qualitative study of Douyin, a short video sharing mobile application in China. Presented at: Social Computing and Social Media Design, Human Behavior and Analytics: 11th International Conference, SCSM 2019, Held as Part of the 21st HCI International Conference, HCII 2019; Jul 26-31, 2019. [doi: 10.1007/978-3-030-21902-4_17]
24. Annual report. : Bilibili Inc; 2024 URL: https://ir.bilibili.com/media/jesfytmo/bilibili-inc-2024-annual-report.pdf [accessed 2025-12-24]
25. Bilibili to change video tracking to viewing duration from number of hits to avoid short clip bias. Yicai Global. 2023 Jun 27. URL: https://www.yicaiglobal.com/news/20230627-bilibili-to-change-video-tracking-to-viewing-duration-from-number-of-hits-to-avoid-short-clip-bias [accessed 2025-12-24]
26. Xiaohongshu (Rednote) after the TikTok refugees' dust has settled. TechBuzz China. 2025. URL: https://techbuzzchina.substack.com/p/xiaohongshu-rednote-after-the-tiktok [accessed 2025-12-24]
27. Hall C, Yu S. China's instagram-like Xiaohongshu making inroads with e-commerce sales. Reuters. 2024. URL: https://www.reuters.com/technology/chinas-instagram-like-xiaohongshu-making-inroads-with-e-commerce-sales-2024-12-16/ [accessed 2025-12-24]
28. Tian Y, Robinson JD. Content analysis of health communication. In: Research Methods in Health Communication: Principles and Application: Routledge; 2014:196-212. [doi: 10.4324/9780203115299-13]
29. Fazeli S, Sabetti J, Ferrari M. Performing qualitative content analysis of video data in social sciences and medicine: the visual-verbal video analysis method. Int J Qual Methods 2023 Oct;22:1-18. [doi: 10.1177/16094069231185452]
30. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. Am J Gastroenterol 2007 Sep;102(9):2070-2077. [doi: 10.1111/j.1572-0241.2007.01325.x] [Medline: 17511753]
31. Zhang Y, Sun Y, Xie B. Quality of health information for consumers on the web: a systematic review of indicators, criteria, tools, and evaluation results. Asso for Info Science & Tech 2015 Oct;66(10):2071-2084. [doi: 10.1002/asi.23311]
32. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999 Feb;53(2):105-111. [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
33. Kaicker J, Borg Debono V, Dang W, Buckley N, Thabane L. Assessment of the quality and variability of health information on chronic pain websites using the DISCERN instrument. BMC Med 2010 Oct 12;8:59. [doi: 10.1186/1741-7015-8-59] [Medline: 20939875]
34. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282. [Medline: 23092060]
35. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol 2013 Apr 29;13:61. [doi: 10.1186/1471-2288-13-61] [Medline: 23627889]
36. Ghosal S, Jain A. Depression and suicide risk detection on social media using fastText embedding and XGBoost classifier. Procedia Comput Sci 2023;218:1631-1639. [doi: 10.1016/j.procs.2023.01.141]
37. Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease prediction by machine learning over big data from healthcare communities. IEEE Access 2017;5:8869-8879. [doi: 10.1109/ACCESS.2017.2694446]
38. Belete DM, Huchaiah MD. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. International Journal of Computers and Applications 2022 Sep 2;44(9):875-886. [doi: 10.1080/1206212X.2021.1974663]
39. Adnan M, Alarood AAS, Uddin MI, Ur Rehman I. Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models. PeerJ Comput Sci 2022;8:e803. [doi: 10.7717/peerj-cs.803] [Medline: 35494796]
40. Prendin F, Pavan J, Cappon G, Del Favero S, Sparacino G, Facchinetti A. The importance of interpreting machine learning models for blood glucose prediction in diabetes: an analysis using SHAP. Sci Rep 2023 Oct 6;13(1):16865. [doi: 10.1038/s41598-023-44155-x] [Medline: 37803177]
41. Wang H, Liang Q, Hancock JT, Khoshgoftaar TM. Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. J Big Data 2024;11(1):45. [doi: 10.1186/s40537-024-00905-w]

42. Yardley L, Spring BJ, Riper H, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
43. Bode L, Vraga EK. See something, say something: correction of global health misinformation on social media. Health Commun 2018 Sep;33(9):1131-1140. [doi: 10.1080/10410236.2017.1331312] [Medline: 28622038]
44. Keelan J, Pavri-Garcia V, Tomlinson G, Wilson K. YouTube as a source of information on immunization: a content analysis. JAMA 2007 Dec 5;298(21):2482-2484. [doi: 10.1001/jama.298.21.2482] [Medline: 18056901]
45. Azak M, Şahin K, Korkmaz N, Yıldız S. YouTube as a source of information about COVID-19 for children: Content quality, reliability, and audience participation analysis. J Pediatr Nurs 2022;62:e32-e38. [doi: 10.1016/j.pedn.2021.06.024] [Medline: 34247879]
46. Wang M, Yao N, Wang J, Chen W, Ouyang Y, Xie C. Bilibili, TikTok, and YouTube as sources of information on gastric cancer: assessment and analysis of the content and quality. BMC Public Health 2024 Jan 2;24(1):57. [doi: 10.1186/s12889-023-17323-x] [Medline: 38166928]
47. Liu H, Peng J, Li L, et al. Assessment of the reliability and quality of breast cancer related videos on TikTok and Bilibili: cross-sectional study in China. Front Public Health 2024;11:1296386. [doi: 10.3389/fpubh.2023.1296386] [Medline: 38317686]
48. Aldous KK, An J, Jansen BJ. View, like, comment, post: analyzing user engagement by topic at 4 levels across 5 social media platforms for 53 news organizations. Presented at: 13th International Conference on Web and Social Media, ICWSM 2019; Jun 11-14, 2019. [doi: 10.1609/icwsm.v13i01.3208]
49. Lim MSC, Wright CJC, Carrotte ER, Pedrana AE. Reach, engagement, and effectiveness: a systematic review of evaluation methodologies used in health promotion via social networking sites. Health Promot J Austr 2016 Feb;27(3):187-197. [doi: 10.1071/HE16057] [Medline: 27719734]
50. Broughton DE, Schelble A, Cipolla K, Cho M, Franasiak J, Omurtag KR. Social media in the REI clinic: what do patients want? J Assist Reprod Genet 2018 Jul;35(7):1259-1263. [doi: 10.1007/s10815-018-1189-2] [Medline: 29766400]
51. Carneiro MM, Koga CN, Mussi MCL, Fradico PF, Ferreira MCF. Quality of information provided by Brazilian Fertility Clinic websites: compliance with Brazilian Medical Council (CFM) and American Society for Reproductive Medicine (ASRM) Guidelines. JBRA Assist Reprod 2023 Jun 22;27(2):169-173. [doi: 10.5935/1518-0557.20220026] [Medline: 35916465]
52. Richardson MA, Park W, Bernstein DN, Mesfin A. Analysis of the quality, reliability, and educational content of youtube videos concerning spine tumors. Int J Spine Surg 2022 Apr;16(2):278-282. [doi: 10.14444/8215] [Medline: 35444036]
53. Scanfeld D, Scanfeld V, Larson EL. Dissemination of health information through social networks: twitter and antibiotics. Am J Infect Control 2010 Apr;38(3):182-188. [doi: 10.1016/j.ajic.2009.11.004] [Medline: 20347636]
54. Murphy D, Balka E, Poureslami I, Leung DE, Nicol AM, Cruz T. Communicating health information: the community engagement model for video production. Can J Commun 2007 Nov 12;32(3-4):383-400. [doi: 10.22230/cjc.2007v32n3a1966]
55. Poecze F, Ebster C, Strauss C. Social media metrics and sentiment analysis to evaluate the effectiveness of social media posts. Procedia Comput Sci 2018;130:660-666. [doi: 10.1016/j.procs.2018.04.117]
56. Tenenboim O. Comments, shares, or likes: what makes news posts engaging in different ways. Soc Media Soc 2022 Oct;8(4):1-15. [doi: 10.1177/20563051221130282]

Abbreviations

DAU: daily active user
GQS: Global Quality Score
IVF: in vitro fertilization
mDISCERN: modified DISCERN
SHAP: Shapley Additive Explanations
XGBoost: Extreme Gradient Boosting

Edited by T Mackey; submitted 10.Sep.2025; peer-reviewed by A Famotire, Y Jiang; revised version received 09.Dec.2025; accepted 09.Dec.2025; published 28.Jan.2026.

Please cite as:
Chu D, Bai X, Guo F
Quality, Reliability, and Dissemination of In Vitro Fertilization–Related Videos on Chinese Social Media: Cross-Sectional Analysis of 300 Short Videos
JMIR Infodemiology 2026;6:e83900
URL: https://infodemiology.jmir.org/2026/1/e83900
doi:10.2196/83900

Review of the Quality and Reliability of Online Arabic Content on Diabetic Retinopathy: Infodemiological Study

Abdullah Ahmed Alabdrabulridha, MBBS; Dalal Mahmoud Alabdulmohsen, MBBS; Maryam Abdullah AlNajjar, MBBS; Ibtisam Ahmed Algouf, MBBS; Abdullah Mohammed Al-Omair, MBBS; Omaima Muneer Alyahya, MBBS; Mesa Ahmed Almahmudi, MBBS; Abdulaziz Ahmed Al Taisan, MBBS

Department of Ophthalmology, College of Medicine, King Faisal University, P.O. 380 Ahsaa, Hofuf, Saudi Arabia

Corresponding Author:
Abdullah Mohammed Al-Omair, MBBS
Department of Ophthalmology, College of Medicine, King Faisal University, P.O. 380 Ahsaa, Hofuf, Saudi Arabia

Abstract

Background: Diabetic retinopathy (DR) is a leading cause of vision loss, particularly in the Middle East. With the rise of online health information, many patients turn to the internet for knowledge about health conditions. However, the accuracy and quality of this information can be questionable, particularly in languages other than English.

Objective: We sought to evaluate the quality and reliability of Arabic websites on DR to address this knowledge gap and improve patient care.

Methods: The first 100 Arabic search results for DR were examined on Google, focusing on patient education websites in Arabic. Content was assessed using a 20-question model, quality was evaluated with the DISCERN instrument, and reliability was measured using the *Journal of the American Medical Association* (*JAMA*) benchmark. Two independent raters conducted evaluations, and data were analyzed with SPSS (IBM Corp). Descriptive statistics were used for website characteristics, and the first 10 Google web pages were compared to others using bivariate analysis with a significance level of $P<.05$.

Results: A Google search yielded 178,000 websites, and the first 100 were examined, with 29 meeting inclusion criteria. Most were hospital or medical center sites (n=20, 69%). The DISCERN assessment showed a low mean score of 36.59(SD 9.32) out of 80 points, with most rated “poor” or “very poor.” The *JAMA* benchmarks indicated low reliability, with 62% (18/29) failing to meet any criteria.

Conclusions: This study identified significant failings in the content, quality, and reliability of Arabic websites on diabetic retinopathy, highlighting the need for stronger evidence-based online resources focused on early disease prevention.

(JMIR Infodemiology 2026;6:e70514) doi:10.2196/70514

KEYWORDS

diabetic retinopathy; diabetes; online health information; arabic content; reliability; internet; online; quality; retinopathy; website; JAMA; DISCERN; health education; Journal of the American Medical Association

Introduction

Diabetic retinopathy (DR) is a major microvascular complication of diabetes mellitus, and it is considered one of the primary causes of permanent vision loss in adults and older individuals across the globe [1]. According to the World Health Organization, the Eastern Mediterranean region has the highest prevalence of diabetes mellitus worldwide. Moreover, it is estimated that DR cases are reaching up to 31% in the Eastern Mediterranean region, which is considered to be higher than the other regions on the globe [2].

Since online-based medical health information has become easily accessible to the population, it facilitates the searching and understanding of disease symptoms, risk factors, and treatment choices. It has been implemented that browsing the internet for health information has become a widespread part of the daily routine of individuals of all ages [3]. Interestingly, in today’s digital age, online search engines are the first resource for almost 75% of patients for medical conditions [4]. People tend to seek medical online information due to the low costs, being less time-consuming, and anonymity; however, online literature can provide mixed results that can be true or misleading, and this can be due to multiple factors, such as the accuracy, readability, and quality of the reported information [5]. Patients are increasingly using the internet to find health-related information that could influence medical decisions; however, there is a risk of encountering commercially influenced content. Hence, these findings indicate that the online information available on DR may not offer sufficient guidance for medical purposes [6].

The quality of online-based ophthalmological diseases was evaluated by several studies. Unfortunately, only a few studies

have documented the quality and reliability of the online content of DR [6]. The studies conducted so far have mainly centered on health-related material available in English, overlooking the need to evaluate the reliability and quality of online information in other languages. It must be mentioned that the main spoken language in the Middle East, which encounters a high rate of DR cases, is Arabic [7,8]. Due to the lack of highly standardized literature on online health content about DR in Arabic, this study seeks to address this gap by conducting a qualified evaluation of Arabic-language websites focusing on DR.

Methods

Study Aim, Design, and Setting

This cross-sectional website analysis was designed to evaluate the reliability and quality of Arabic online information about DR. Google.com, the most widely used search engine worldwide, held a 90.29% market share in Asia and a 97.19% market share in Africa as of May 2024. Arabic-speaking countries in the Middle East span both continents. For example, Saudi Arabia has a market share of 95.60% while Egypt has 97.38% [9]. The engine was used on May 1, 2024, to search for the Arabic term for DR, “ ,” in incognito mode using a new account to avoid browser bias. The first 100 search results were examined, simulating a patient’s or a general reader’s search behavior.

Inclusion and Exclusion Criteria

Websites were included if they were written primarily in Arabic and focused on providing educational content about DR for patients or the general public. Eligible websites were required to contain written text that addressed DR-specific topics, such as causes, symptoms, diagnosis, treatment options, and preventive strategies. The inclusion was limited to the first 100 search results to reflect typical patient behavior when searching for online health information.

Websites were excluded if they were not written in Arabic, targeted health care professionals (such as academic articles or clinical guidelines), or consisted solely of multimedia content like videos or audio recordings without accompanying text. In addition, websites that required login credentials, subscription access, or were otherwise inaccessible were excluded. Duplicate URLs among the first 100 search results were also removed. Finally, websites with irrelevant content, such as general diabetes pages lacking specific focus on DR, social media posts, news articles, blogs, forums, or advertisements, were excluded. The searching process is further illustrated in Figure 1.

Figure 1. Flowchart of the search process and results.

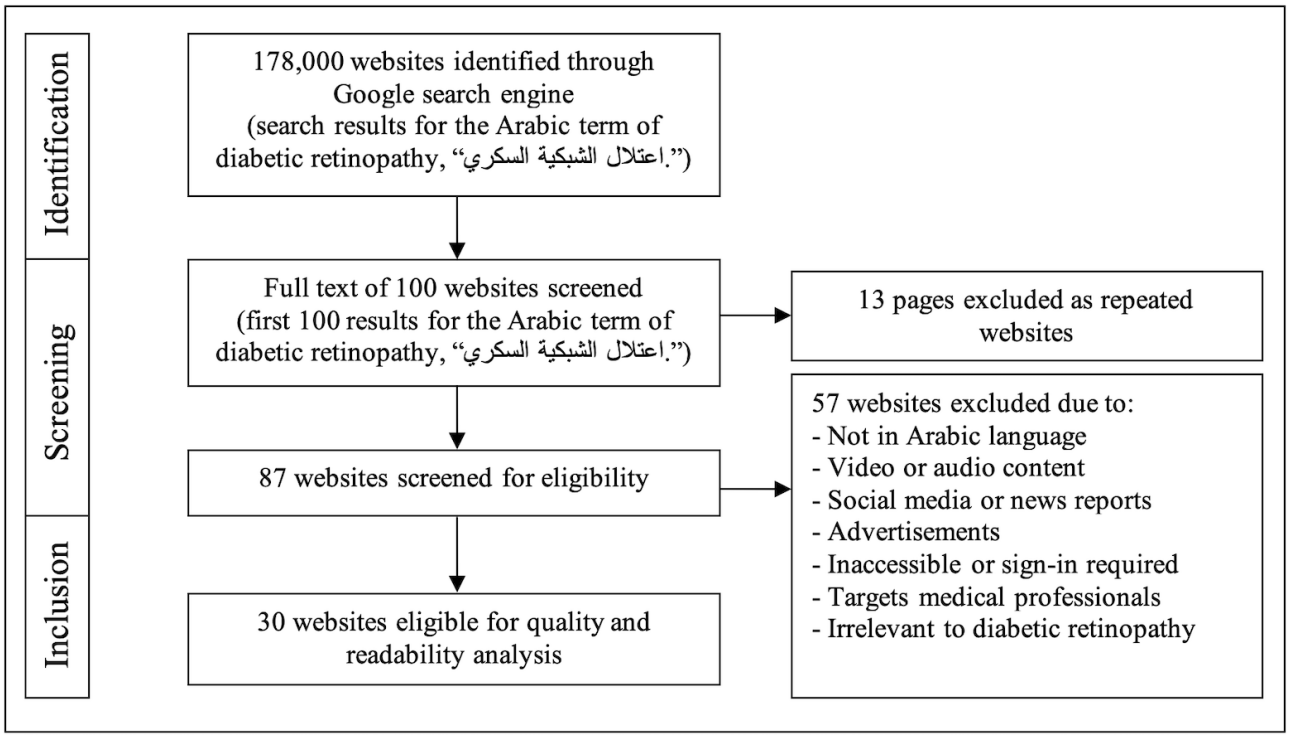


Table 1 exhibits the list of websites eligible for evaluation. Some websites, such as Mayo Clinic, have identical URLs in different languages but automatically localize content to Arabic. Websites were classified into four main categories: hospitals and medical centers, health portals (websites dedicated to health information), commercial (websites selling a service or product), and nonprofit organizations.

Table . Websites eligible for evaluation.

Website type	Website name	URL
Nonprofit organization	Mayo Clinic [10]	https://www.mayoclinic.org
Health portals	Webteb [11]	https://www.webteb.com
Health portals	MSD Manuals [12]	https://www.msdmanuals.com
Commercial	Altibbi [13]	https://altibbi.com
Hospitals and medical centers	Cleveland Clinic Abu Dhabi [14] https://www.msdmanuals.com	https://www.clevelandclinicabudhabi.ae
Hospitals and medical centers	Institut Català De Retina [15]	https://icrcat.com
Hospitals and medical centers	Barraquer UAE Eye Hospital [16]	https://www.barraquer.com
Hospitals and medical centers	Dünyagöz [17]	https://www.dunyagoz.com
Hospitals and medical centers	Bangkok Hospital [18]	https://www.bangkokhospital.com
Hospitals and medical centers	Med Care [19]	https://www.medcare.ae
Hospitals and medical centers	Northwest Eye Surgeons [20]	https://www.nweyes.com
Hospitals and medical centers	Dr. Haifa Eye Hospital [21]	https://www.drhaifaeyehospital.com
Hospitals and medical centers	King Khaled Eye Specialist Hospital [22]	https://pep.kkesh.med.sa
Hospitals and medical centers	Royal Spanish Center [23]	https://www.royalspanishcenter.com
Hospitals and medical centers	Jgemc [24]	http://www.jgemc.com
Commercial	Vezeeta [25]	https://www.vezeeta.com
Commercial	Ilajak [26]	https://www.ilajak.com
Hospitals and medical centers	Dr Mahmoud Hassaan [27]	https://www.drmahmoud-hassaan.com
Hospitals and medical centers	Eye City Center [28]	https://www.eyecitycenter.com
Nonprofit organization	Gulf Health Council [29]	https://www.ghc.sa
Health portals	Elconsolto [30]	https://www.elconsolto.com
Hospitals and medical centers	Alkahhal [31]	https://alkahhal.com.sa
Hospitals and medical centers	Ebsaar [32]	https://ebsaar.com
Hospitals and medical centers	Smart Laser Eye Center [33]	https://www.smartlasereyecenter.com
Hospitals and medical centers	Jordan Finland Modern Hospital [34]	https://jfmhospital.com
Hospitals and medical centers	Novomed [35]	https://www.novomed.com
Commercial	Tebcan [36]	https://tebcan.com
Hospitals and medical centers	Andalusia Clinic [37]	https://www.andalusiaclinic.com
Hospitals and medical centers	Dr-Oyoun [38]	https://dr-oyoun.com
Hospitals and medical centers	Magrabi Hospital [39]	https://www.magrabi.com

Assessment Tools

Content Assessment

To describe the content characteristics of the assessed websites, 20 questions were adapted from an established model by Kloosterboer et al [6] specific to DR websites. We used the model's questions to create a scoring system ranging from 0 (no content criteria fulfilled) to 100 (all content questions covered). Of the original 26 questions, 20 were selected for their relevance to the Arabic language content and patient-oriented information. Questions that were not directly applicable were excluded.

Quality Assessment

The DISCERN instrument, a well-known tool for evaluating patient-targeted content quality, was also used. This questionnaire contains 16 questions with clear instructions for objective assessment, with answers ranging from 1 (no sufficient answer) to 5 (efficiently answered) [40]. The total score ranges from 16 to 80 total points. The score categories were defined by Novin et al [40], who used this tool to assess DR online information for patients, to be as follows: "Excellent (75 - 63 points), Good (62 - 51 points), Average (50 - 39 points), Poor (38 - 28 points), and Very Poor (<28 points)."

Reliability Assessment

The *Journal of the American Medical Association* (*JAMA*) benchmark, used to evaluate website reliability, has 4 standards, namely authorship, attribution, currency, and disclosure. Each website receives 1 point when it fulfills a standard's criteria, with a maximum total score of 4 points. Websites scoring 3 or higher are considered highly reliable, while those scoring lower are considered low reliability [41].

Evaluation Process

The evaluation was performed by 2 independent raters (Evaluators A and B), followed by a shared revision session to determine a final rating and resolve any disagreements from the initial evaluation. The 2 raters were senior medical students trained in the use of the DISCERN and *JAMA* benchmarks, following a structured protocol to ensure consistency and objectivity. While not board-certified ophthalmologists, the raters applied the tools according to standardized instructions that do not require clinical expertise. Raters followed the instructions of the tools to objectively determine appropriate scores for each website. The final rating was derived from the initial independent evaluations and did not substantially deviate from either rater's original score. In all instances, the final rating either corresponded to one of the initial assessments or reflected a minor adjustment reached through consensus during the revision session.

Data Management and Analysis

All data was organized in an Excel (Microsoft) spreadsheet and imported into IBM SPSS (version 21; IBM Corp). Descriptive statistics, such as means and SDs, were used to summarize website characteristics and evaluation scores. Normality tests were conducted to determine the data distribution (normal or nonparametric), and appropriate statistical tests were applied. Because internet users are more likely to view websites on the first page of Google (the first 10 websites) [42], these were considered to be the most viewed and therefore grouped together to be compared with the websites of other pages using both the Mann-Whitney *U* test and independent 2-tailed *t* test according to the distribution of the data. The direction of association between the DISCERN and *JAMA* scores was examined using nonparametric measures, more specifically Spearman rank correlation. A *P* value below .05 was considered statistically significant. This study was reported in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [43].

Results

Website Selection Process and Overview

Our search for the Arabic term for DR using the Google search engine yielded 178,000 websites. Examining the first 100 revealed 13 duplicates, which were excluded, and further filtering of the remaining 87 for eligibility led to the exclusion of pages in languages other than Arabic, video or audio content only, inaccessible sites requiring visitors to sign in, those targeting medical professionals only, irrelevant topics, advertisements, social media posts, and news reports. Among the 30 pages that were eligible for further analysis, one more website was excluded and removed from the dataset as it became unavailable during the analysis process.

Of the 29 websites finally included, the majority (n=20, 69%) were classified as hospitals and medical centers. Commercial pages made up 4 (13.8%) websites. Around 3 (10.3%) websites fell under the health portal classification. The remaining 2 (6.9%) websites belonged to the foundation or nonprofit organization category.

Quality Assessment by DISCERN

The overall mean score of DISCERN for all websites was low (mean 36.59,SD 9.322), ranging from 17 to 61 points. None of the websites reached the "excellent" category (≥63 points). Only 2 of 29 (6.9%) websites scored in the "good" range (≥51 points). A total of 7 of 29 (24.1%) pages fell into the "average" category with a score ranging from 39 to 50 points. The majority of the included websites (16/29, 55.2%) had a "poor" score (38 - 28 points), and the remaining 4 of 29 (13.8%) pages achieved a "very poor" score of less than 28 points.

The overall quality score (DISCERN item 16) of the identified 29 websites had an average of 2.45 (SD 0.910). The DISCERN items with the highest scores were related to the relevancy of the content, the variety of treatment options, and how each treatment works (mean 4.59, SD 0.983; mean 4.00, SD 1.000; and mean 4.10, SD 1.263, respectively; Table 2).

Table . Average scores for each DISCERN item in evaluating 29 websites, ranging from 1 to 5.

Item	Score
1. Are the aims clear?	1.38
2. Does it achieve its aims?	1.55
3. Is it relevant?	4.59
4. Is it clear what sources of information were used to compile the publication (other than the author or producer)?	1.45
5. Is it clear when the information used or reported in the publication was produced?	1.48
6. Is it balanced and unbiased?	3.62
7. Does it provide details of additional sources of support and information?	1.66
8. Does it refer to areas of uncertainty?	1.10
9. Does it describe how each treatment works?	4.10
10. Does it describe the benefits of each treatment?	2.90
11. Does it describe the risks of each treatment?	1.90
12. Does it describe what would happen if no treatment is used?	1.69
13. Does it describe how the treatment choices affect overall quality of life?	1.17
14. Is it clear that there may be more than one possible treatment choice?	4.00
15. Does it provide support for shared decision-making?	1.55
16. Based on the answers to all of the above questions, rate the overall quality of the publication as a source of information about treatment choices	2.45

Several websites scored low in relation to revealing the explicit aims of their content and their source of information (mean 1.38,SD 1.015 and mean 1.45, SD 1.055, respectively). In addition, not many pages reported the date their content was created or published (mean score 1.48, SD 0.949). Few websites mentioned the possible risks of each treatment (mean score 1.90, SD 1.345), and almost none of the websites (1/29, 3.4%) highlighted any “gray” areas or uncertainties about the outcomes of the treatments (mean score 1.10, SD 0.557). Nearly all websites (27/29, 93.1%) scored 1 point out of 5 in the question related to the effect of treatment on quality of life (mean 1.17, SD 0.658; Table 2).

Reliability Assessment by *JAMA* Benchmarks

None of the websites obtained a score higher than 2 on the *JAMA* benchmarks tool (mean 0.45, SD 0.632), placing them all in the “low reliability” category. Over two-thirds (18/29, 62.1%) of the websites failed to meet any criteria, while the remaining websites (11/29, 37.9%) met only 1 or 2. The most common criterion met was currency, with 10 of 29 (34.5%) websites complying with it and an average score of 0.34 (SD 0.484). Authorship (mean 0.03, SD 0.186) and disclosure (mean 0.07, SD 0.258) were only displayed on 1 and 2 web pages (1/29, 3.4% and 2/29, 6.9%, respectively). No website has fulfilled the attribution benchmark. A very weak positive correlation between the DISCERN and *JAMA* scores was observed (ρ=0.130; *P*=.50).

Content Assessment

Out of 20 questions, a single website (1/29, 3.4%) answered the most, covering 75% of the questions. On the other hand, (2/29, 6.9% of the group) websites answered the least number of questions, only covering 15%. While most websites (27/29, 93.1%) explained what DR was and how it was treated, very few mentioned the screening period for DR (8/29, 27.6%). Only 2 of 29 (6.9%) websites discussed the reversibility of vision loss caused by DR. Moreover, 5 of 29 (17.2%) of the included websites covered the surgical options to treat DR and its possible risks. The vast majority (25/29, 86.2%) lacked images of DR (Table 3).

Table . Twenty questions about the content related to diabetic retinopathy with frequencies and percentage.

Questions	Yes, n (%)	No, n (%)
What is diabetic retinopathy?	27 (93.1)	2 (6.9)
What are the symptoms of diabetic retinopathy?	22 (75.9)	7 (24.1)
What is the difference between nonproliferative and proliferative diabetic retinopathy?	18 (62.1)	11 (37.9)
How is diabetic retinopathy diagnosed?	19 (65.5)	10 (34.5)
When should screening start?	8 (27.6)	21 (72.4)
What are the risk factors for diabetic retinopathy?	19 (65.5)	10 (34.5)
Can anything be done to reverse diabetic retinopathy?	9 (31)	20 (69)
What percentage of patients become legally blind from diabetic retinopathy?	2 (6.9)	27 (93.1)
How can vision loss be prevented?	11 (37.9)	18 (62.1)
Is vision loss reversible?	2 (6.9)	27 (93.1)
How is diabetic retinopathy treated?	27 (93.1)	2 (6.9)
What is panretinal photocoagulation, and what are the complications associated with it?	8 (27.6)	21 (72.4)
What is an anti-VEGF[a] injection and what are the complications associated with anti-VEGF therapy?	6 (20.7)	23 (79.3)
Are anti-VEGF injections or laser a cure or do they need to be repeated?	8 (27.6)	21 (72.4)
What are the surgical treatments for diabetic retinopathy and what are the potential complications?	5 (17.2)	24 (82.8)
What is tractional retinal detachment?	12 (41.4)	17 (58.6)
What is diabetic macular edema?	12 (41.4)	17 (58.6)
Are there any oral medications that can alter the progression of diabetic retinopathy?	1 (3.4)	28 (96.6)
Which age group is most commonly affected by diabetic retinopathy?	2 (6.9)	27 (93.1)
Does the source show pictures of diabetic retinopathy?	4 (13.8)	25 (86.2)

[a]VEGF: vascular endothelial growth factor.

Comparison of the Websites on the First Page and Additional Pages

The Mann-Whitney *U* test revealed no significant difference in *JAMA* scores between the first 10 websites (displayed on the first page) and those on subsequent pages (1-tailed *P*=.12). Similarly, bivariate analysis showed no significant difference in the DISCERN and content scores between the websites of the first page and other pages (*P*=.72). Detailed individual website scores for quality, reliability, and content are presented in Table 4.

Table . Included websites' scores on quality, reliability, and content.

Website name	Quality assessment by DISCERN (16-80)	Class	Reliability assessment by *JAMA*[a] (0-4)	Class	Content score (out of 20)	Content score (out of 100)
Mayo Clinic	55	Good	2	Low reliability	15	75
Webteb	37	Poor	1	Low reliability	11	55
MSD Manuals	38	Poor	2	Low reliability	9	45
Altibbi	44	Average	0	Low reliability	9	45
Cleveland Clinic Abu Dhabi	37	Poor	0	Low reliability	10	50
Institut Català De Retina	29	Poor	1	Low reliability	4	20
Barraquer Uae Eye Hospital	25	Very poor	0	Low reliability	7	35
Dünyagöz	17	Very poor	1	Low reliability	3	15
Bangkok Hospital	35	Poor	0	Low reliability	9	45
Med Care	35	Poor	0	Low reliability	9	45
Northwest Eye Surgeons	22	Very poor	0	Low reliability	3	15
Dr. Haifa Eye Hospital	38	Poor	0	Low reliability	10	50
King Khaled Eye Specialist Hospital	36	Poor	0	Low reliability	11	55
Royal Spanish Center	41	Average	0	Low reliability	13	65
Jgemc	30	Poor	0	Low reliability	5	25
Vezeeta	27	Very poor	0	Low reliability	6	30
Ilajak	41	Average	0	Low reliability	6	30
Dr Mahmoud Hassaan	41	Average	1	Low reliability	7	35
Eye City Center	48	Average	1	Low reliability	7	35
Gulf Health Council	29	Poor	0	Low reliability	5	25
Elconsolto	61	Good	0	Low reliability	5	25
Alkahhal	41	Average	0	Low reliability	12	60
Ebsaar	28	Poor	0	Low reliability	8	40
Smart Laser Eye Center	37	Poor	1	Low reliability	8	40
Jordan Finland Hospital	34	Poor	1	Low reliability	8	40
Novomed	35	Poor	1	Low reliability	5	25
Tebcan	37	Poor	1	Low reliability	5	25
Andalusia Clinic	34	Poor	0	Low reliability	6	30
Dr-Oyoun	49	Average	0	Low reliability	6	30

[a]*JAMA*: *Journal of the American Medical Association.*

Discussion

Principal Findings

Our study assessed the reliability and quality of the contents of 29 Arabic websites covering DR based on DISCERN and *JAMA* benchmarks. Using the DISCERN tool to evaluate the quality of each website has shown an overall poor quality of all websites evaluated, with an average score of 36.59 (SD 9.322). None of the 29 websites assessed achieved "excellent," with the majority of the websites falling under the "poor" (16/29, 55.2%) and "very poor" (4/29, 13.8%) categories. The content of the websites seems to falter in the first section of the DISCERN instrument focusing on the trustworthiness of the content presented (questions 1 - 8), which reflects the unreliability of the websites offering information on DR in Arabic. In addition, comprehensive reporting of treatment options with their associated benefits and risks was inconsistent between the websites, compromising the overall quality of the content presented. Our study found that the average score of the overall quality on the DISCERN tool is 2.45 (SD 0.9), which falls under the category of having "potentially important but not serious shortcomings." This outcome is similar to the results found by Novin et al [40], who evaluated the information about DR on US-based online websites and observed a mean score of 2.09 (SD 0.594), and similar to our results, no website was classified as "excellent" in their analysis. Moreover, both studies observed a deficiency in directing the readers to discuss their conditions with their physicians. While a perfect score of 5 signifies the highest quality, none of the analyzed websites achieved this benchmark.

Given the consistently poor quality and reliability scores across websites, these findings may be explained by several underlying factors. Hypotheses include (1) the absence of standardized quality controls in Arabic web content, (2) limited contributions from medical institutions in developing and maintaining patient education resources, and (3) a general lack of regulatory oversight governing the accuracy and reliability of health information published online. Such systemic gaps likely underlie the consistently low quality and reliability scores observed in our assessment.

Similar to the DISCERN assessment of the websites, assessing the 29 websites included with the *JAMA* benchmarks tool has revealed that all the websites are of low reliability. None of the websites evaluated obtained a score higher than 2, with an average score of 0.45 (SD 0.632). Comparing our results to the study done on dry eye disease websites, their assessment using *JAMA* has found that all the websites failed to meet half the benchmarks as well, with an average of 1.9 (SD 0.1) [44]. The low average in our study is due to more than two-thirds (18/29, 62.1%) of the websites assessed failing to meet any of the 4 *JAMA* benchmark criteria for a reliable website. A study conducted by Kloosterboer et al [6] on assessing online information regarding DR has also found that none of the websites evaluated has achieved all 4 *JAMA* benchmarks.

Upon evaluating each website's content, it was observed that no website has covered all questions, with the highest-scoring website only fulfilling 75% of the questions (n=15). While most websites included information on what DR is and how it is treated, many of them (20/29, 69%) lacked information about when screening should begin. The study by Novin et al [6] also noted this, pointing out that there was insufficient content on the DR screening intervals for type 1 and type 2 diabetes. A substantial number of the evaluated websites did not thoroughly address every therapeutic choice with its correlated risks and benefits. This was also a common finding in another study assessing DR-related internet resources for patients, where most websites scored in the lower range in questions related to treatment options.

The study found the average score for the balance and unbiasedness item of the DISCERN instrument to be 3.62 (SD 1.24) for Arabic websites covering DR. This rating may be attributed to the inclusion of commercial websites in the evaluation, despite excluding those explicitly labeled as "advertisement" or "ad" in Google search results. These commercial sites, while offering health services or products, also provide health educational content. The presence of commercial links and advertisements can hinder users' ability to locate reliable information, contributing to a "misinfodemic" and making it challenging for patients to access trustworthy health resources [45].

We believe that the gap found in the quality of content reported on all the assessed websites reflects the current health care practices in the region, where collaborative decision-making and thorough explanation of treatment risks and benefits may not always be given priority. A systematic literature review of patient-centered care (PCC) in the Middle East concluded that while there is support for adopting PCC in the Middle East and North African region, its implementation is still limited [46]. Webair identified barriers to adopting a PCC approach at multiple levels, mostly related to communication, suggesting a preference for a physician-driven approach that may not place as much emphasis on discussing treatment alternatives with patients [47].

Our study has its own set of limitations that require addressing for future investigation and research. First, Google was the primary search engine used in this study. Although it is a well-renowned and widely used search engine, the use of other search engines could provide more websites that were not taken into consideration by relying solely on Google [9]. The second limitation is the exclusive focus on written Arabic-language content. Video-based or multimedia educational resources, which are increasingly used by patients, were excluded from this analysis. Evaluating such content could provide further insights into the quality of health information consumed by the general population through popular platforms, such as WhatsApp (Meta Platforms Inc), YouTube (Google), or TikTok (ByteDance). While the quality of the content was assessed thoroughly in our study, assessing the readability of the websites included could provide insight into whether the information provided can be adequately comprehended by patients or not. Further studies should be performed in this area to assess whether websites are within the acceptable reading levels of online educational materials.

The study shows the need to improve the quality and reliability of Arabic-language online content on DR. Content creators should follow best practices in health communication, including clear authorship, proper source attribution, transparency regarding sponsorship, and regular content updates. Information should also be patient-centered, discussing treatment options, associated risks, and preventive measures like regular screening.

At the policy level, national guidelines are needed to ensure the quality of Arabic online health information. Policymakers may consider implementing standardized accreditation systems like the Health on the Net Code of Conduct, adapted to the region's linguistic and cultural context, to combat misinformation and enhance public health literacy.

Conclusion

Our study's findings disclose that Arabic-language websites providing information on DR treatment are significantly deficient in quality, reliability, and content. The DISCERN assessment tool displayed a "poor" score for most of the analyzed websites (16/29, 55.2%). In addition, according to the *JAMA* benchmark criteria, all of the websites showed low reliability, and none of them met the attribution benchmark. These findings highlight the necessity of enhancing the Arabic-language websites discussing DR treatment due to the high prevalence of this condition among patients in the Middle East and the insufficient high-quality online resources available. Middle Eastern health care organizations should collaborate to provide reliable, evidence-based online resources addressing this serious condition. While most websites discuss the treatment of DR, more information is needed on the associated risks and benefits, as well as a stronger focus on joint decision-making. In addition, Arabic websites should encourage screening and preventive measures to enhance patient outcomes.

Acknowledgments

The authors would like to express deep appreciation for all those who willingly contributed their time to this research.

Funding

We acknowledge the financial support for this research provided by the Deanship of Scientific Research at King Faisal University in Al-Ahsa, Kingdom of Saudi Arabia (grant number KFU241612).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: AAA, DMA, MAA, IAA, AMA, OMA
Methodology: AAA, DMA, MAA, IAA, AMA, OMA
Investigation: AAA, DMA, MAA, IAA, AMA, OMA
Software: AAA, DMA, MAA, IAA, AMA, OMA
Writing–original draft: AAA, DMA, MAA, IAA, AMA, OMA
Writing–review & editing: AAAT
Supervision: AAAT
Project administration: AAAT
Funding acquisition: AAAT

Conflicts of Interest

None declared.

References

1. Chaudhari YS, Matkar SS, Bhujbal SS, Walunj VA, Bhor NS, Vyavhare RD. Diabetes mellitus: a review. Int J Adv Res Sci Commun Technol 2023 Mar 1:16-22. [doi: 10.48175/IJARSCT-8551]
2. Heiran A, Azarchehry SP, Dehghankhalili S, Afarid M, Shaabani S, Mirahmadizadeh A. Prevalence of diabetic retinopathy in the Eastern Mediterranean Region: a systematic review and meta-analysis. J Int Med Res 2022 Oct;50(10):3000605221117134. [doi: 10.1177/03000605221117134] [Medline: 36314851]
3. Zhao YC, Zhao M, Song S. Online health information seeking behaviors among older adults: systematic scoping review. J Med Internet Res 2022 Feb 16;24(2):e34790. [doi: 10.2196/34790] [Medline: 35171099]
4. Finney Rutten LJ, Blake KD, Greenberg-Worisek AJ, Allen SV, Moser RP, Hesse BW. Online health information seeking among us adults: measuring progress toward a healthy people 2020 Objective. Public Health Rep 2019;134(6):617-625. [doi: 10.1177/0033354919874074] [Medline: 31513756]
5. Singh S, Saini R, Sagar R. Quality of web-based information on attention deficit hyperactivity disorder. Asian J Psychiatr 2022 May;71:103071. [doi: 10.1016/j.ajp.2022.103071] [Medline: 35303590]

6. Kloosterboer A, Yannuzzi NA, Patel NA, Kuriyan AE, Sridhar J. Assessment of the quality, content, and readability of freely available online information for patients regarding diabetic retinopathy. JAMA Ophthalmol 2019 Nov 1;137(11):1240-1245. [doi: 10.1001/jamaophthalmol.2019.3116] [Medline: 31436789]
7. Abdulaziz Al Dawish M, Alwin Robert A, Braham R, et al. Diabetes mellitus in Saudi Arabia: a review of the recent literature. Curr Diabetes Rev 2016 Oct 26;12(4):359-368. [doi: 10.2174/1573399811666150724095130]
8. Albadrani MS, Alrehaili AM, Alahmadi SH, et al. Awareness of diabetic retinopathy among patients with type 2 diabetes mellitus in primary healthcare centers in Madinah, Saudi Arabia: a cross-sectional study. Cureus 2023 Nov;15(11):e49718. [doi: 10.7759/cureus.49718] [Medline: 38033448]
9. Browser market share worldwide. StatCounter Global Stats. URL: https://gs.statcounter.com [accessed 2024-06-28]
10. Diabetic retinopathy. Mayo Clinic. URL: https://www.mayoclinic.org/diseases-conditions/diabetic-retinopathy/symptoms-causes/syc-20371611 [accessed 2025-10-08]
11. Diabetic retinopathy. WebTeb. URL: https://www.webteb.com/diabetes/diseases/ - - [accessed 2025-10-08]
12. Mehta S. Diabetic retinopathy. MSD Manuals. URL: https://www.msdmanuals.com/home/eye-disorders/retinal-disorders/diabetic-retinopathy [accessed 2025-10-08]
13. Diabetic retinopathy. Altibbi. URL: https://altibbi.com/ - / - / - - [accessed 2025-10-08]
14. Diabetic retinopathy. Cleveland Clinic Abu Dhabi. URL: https://www.clevelandclinicabudhabi.ae/ar-ae/health-hub/health-resource/diseases-and-conditions/diabetic-retinopathy [accessed 2025-10-08]
15. Diabetic retinopathy. Institut Català de Retina. URL: https://icrcat.com/ar/ - - / - - / [accessed 2025-10-08]
16. Diabetic retinopathy. Barraquer UAE Eye Hospital. URL: https://www.barraquer.com/ar-uae/pathology/tll-lshbky-lskwry [accessed 2025-10-08]
17. Diabetic retinopathy. Dünyagöz. URL: https://www.dunyagoz.com/ar/medical-units/retina/diabetic-retinopathy [accessed 2025-10-08]
18. Do not overlook diabetic retinopathy as it can lead to complete vision loss!. Bangkok Hospital. URL: https://www.bangkokhospital.com/ar/bangkok/content/diabetic-retinopathy [accessed 2025-10-08]
19. Diabetic retinopathy. Medcare. URL: https://www.medcare.ae/ar/services/view/ophthalmology/diabetic-retinopathy.html [accessed 2025-10-08]
20. Diabetic retinopathy. Northwest Eye Surgeons. URL: https://web.archive.org/web/20231110015353mp_/https://www.nweyes.com/retina-vitreous-seattle/diabetic-retinopathy/ [accessed 2025-10-08]
21. Diabetic retinopathy. Dr Haifa Eye Hospital. URL: https://www.drhaifaeyehospital.com/ar/services/diabetic-retinopathy [accessed 2025-10-08]
22. Diabetic retinopathy. King Khaled Eye Specialist Hospital. URL: https://pep.kkesh.med.sa/?p=1446 [accessed 2025-10-08]
23. Learn about diabetic retinopathy: its causes and treatment methods. Royal Spanish Center. URL: https://royalspanishcenter.com/ - - - - - - /?lang=ar [accessed 2025-10-08]
24. Diabetic retinopathy. Jordan German Eye Center. URL: https://web.archive.org/web/20230321201534/http://www.jgemc.com/index.php/ar/consultancy-clinics/2016-07-23-22-08-43/8-department/54-diabetic-retinopathy [accessed 2025-10-08]
25. Diabetic retinopathy. Vezeeta. URL: https://www.vezeeta.com/ar/ - / / - - [accessed 2025-10-08]
26. Diabetic retinopathy: a comprehensive guide to symptoms, causes, and prevention. Ilajak Medical. URL: https://www.ilajak.com/ar/blog/diabetic-retinopathy [accessed 2025-10-08]
27. Prevention of diabetic retinopathy in 4 steps. Dr Mahmoud Hassaan. URL: https://www.drmahmoud-hassaan.com/blog/ - - - - - [accessed 2025-10-08]
28. Diabetic retinopathy. Al Sharq City Eye Center. URL: https://www.eyecitycenter.com/ar/blogs/23/aaatlal-alshbky-alskry [accessed 2025-10-08]
29. Diabetic retinopathy. Gulf Health Council. URL: https://yourhealthguide.ghc.sa/subjects/ - - -diabetic-retinopathy/ [accessed 2025-10-08]
30. Najah S. Neglecting it leads to blindness — 3 ways to treat diabetic retinopathy. Elconsolto. URL: https://www.elconsolto.com/eye-clinic/eye-clinic-news/details/2022/10/30/2315005/ - - -3- - - - - [accessed 2025-10-08]
31. Learn about how diabetic retinopathy. Alkahhal. URL: https://alkahhal.com.sa/ar/advices/ - - [accessed 2025-10-08]
32. Complications of diabetes. Ebsaar. URL: https://ebsaar.com/ar/services/diabetes-complications/ [accessed 2025-10-08]
33. Diabetic retinopathy. Smart Laser Eye Center. URL: https://www.smartlasereyecenter.com/ar/diabetic-retinopathy-disease/ [accessed 2025-10-08]
34. Diabetic retinopathy. Jordan Finland Modern Hospital. URL: https://web.archive.org/web/20240617224414/https://jfmhospital.com/ - - -diabetic-retinopathy/ [accessed 2025-10-08]
35. Diabetic retinopathy treatment in Dubai. Novomed. URL: https://www.novomed.com/services/specialized-clinics/ophthalmology/diabetic-retinopathy/ [accessed 2025-10-08]
36. Eye diseases: diabetic retinopathy. Tebcan. URL: https://tebcan.com/ar/Jordan/tebline/Admin/ /Articles/ - - - - - 2041 [accessed 2025-10-08]
37. Diabetic retinopathy. Andalusia Clinic. URL: https://andalusiaksa.com/ - - [accessed 2025-10-08]
38. Diabetic retinopathy. Dr-Oyoun. URL: https://dr-oyoun.com/ - - / [accessed 2025-10-08]

39. Diabetic retinopathy. Magrabi Health. URL: https://www.magrabihealth.com/blog/diabetic-retinopathy [accessed 2025-10-08]
40. Novin S, Konda SM, Xing B, Bange M, Blodi B, Burckhard B. Diabetic retinopathy online: a powerful opportunity for revision. J Consum Health Internet 2020 Jul 2;24(3):251-268. [doi: 10.1080/15398285.2020.1791668]
41. Ozduran E, Büyükçoban S. Evaluating the readability, quality and reliability of online patient education materials on post-COVID pain. PeerJ 2022;10:e13686. [doi: 10.7717/peerj.13686] [Medline: 35880220]
42. Eysenbach G, Köhler C. How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. BMJ 2002 Mar 9;324(7337):573-577. [doi: 10.1136/bmj.324.7337.573] [Medline: 11884321]
43. Elm EV, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007 Oct 20;335(7624):806-808. [doi: 10.1136/bmj.39335.541782.AD]
44. Oydanich M, Kuklinski E, Asbell PA. Assessing the quality, reliability, and readability of online information on dry eye disease. Cornea 2022 Aug 1;41(8):1023-1028. [doi: 10.1097/ICO.0000000000003034] [Medline: 35344972]
45. Cai HC, King LE, Dwyer JT. Using the Google search engine for health information: is there a problem? Case study: supplements for cancer. Curr Dev Nutr 2021 Feb;5(2):nzab002. [doi: 10.1093/cdn/nzab002] [Medline: 33937613]
46. Alkhaibari RA, Smith-Merry J, Forsyth R, Raymundo GM. Patient-centered care in the Middle East and North African region: a systematic literature review. BMC Health Serv Res 2023 Feb 9;23(1):135. [doi: 10.1186/s12913-023-09132-0] [Medline: 36759898]
47. Laher I, editor. Handbook of Healthcare in the Arab World: Springer Nature; 2021. [doi: 10.1007/978-3-030-36811-1]

Abbreviations

DR: diabetic retinopathy
***JAMA*:** *Journal of American Medical Association*
PCC: patient-centered care
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by R Cuomo; submitted 23.Dec.2024; peer-reviewed by A Jafarizadeh, G Lim; revised version received 31.Aug.2025; accepted 16.Sep.2025; published 07.Jan.2026.

Please cite as:
Alabdrabulridha AA, Alabdulmohsen DM, AlNajjar MA, Algouf IA, Al-Omair AM, Alyahya OM, Almahmudi MA, Al Taisan AA
Review of the Quality and Reliability of Online Arabic Content on Diabetic Retinopathy: Infodemiological Study
JMIR Infodemiology 2026;6:e70514
URL: https://infodemiology.jmir.org/2026/1/e70514
doi:10.2196/70514

Original Paper

Leveraging AI for Analysis of Digital Health Information on Cancer Prevention Among Arab Youth and Adults: Content Analysis

Alia Komsany[1], PhD; Obada Al Zoubi[2], PhD; Laetitia Sebaaly[3], MA; Gabrielle Harrison[3], BA; Orysya Soroka[1], MS; Safa ElKefi[4], PhD; David Scales[1], MD, PhD; Erica Phillips[1], MS, MD; Laura C Pinheiro[1], MPH, PhD; Israa Ismail[1], BA; Perla Chebli[5], MPH, PhD

[1]Division of General Internal Medicine, Weill Cornell Medicine, New York City, NY, United States

[2]Independent Researcher, Boston, MA, United States

[3]Independent Consultant, New York, NY, United States

[4]School of System Sciences and Industrial Engineering, Watson College of Engineering, Binghamton University, New York, NY, United States

[5]NYU Langone Health, New York City, NY, United States

Corresponding Author:
Alia Komsany, PhD
Division of General Internal Medicine
Weill Cornell Medicine
Division of General Internal Medicine, 530 East 70th Street
New York City, NY, 10021
United States
Phone: 1 6469625036
Email: alk4019@med.cornell.edu

Abstract

Background: As TikTok (ByteDance) grows as a major platform for health information, the quality and accuracy of Arabic-language cancer prevention content remain unknown. Limited access to culturally relevant and evidence-based information may exacerbate disparities in cancer knowledge and prevention behaviors. Although large language models offer scalable approaches for analyzing online health content, their utility for short-form video data, especially in underrepresented languages, has not been well established.

Objective: We aimed to characterize and evaluate the quality of Arabic-language TikTok videos on cancer prevention and explore the use of large language models for scalable content analysis.

Methods: We used the TikTok research application programming interface and a GPT-assisted keyword strategy to collect Arabic-language TikTok videos (2021-2024). From an initial collection of 1800 TikTok videos, 320 were eligible after preprocessing. Of these, the top 25% (N=30) most-viewed were analyzed and manually coded for content type, cancer type, uploader identity, tone and register, scientific citation, and disclaimers. Video quality was assessed using the Patient Education Materials Assessment Tool for Audiovisual Materials for understandability and actionability, and the Global Quality Scale (GQS). GPT-4 was used to generate artificial intelligence annotations, which were compared to human coding for select variables.

Results: The top 25% (N=30) most-viewed videos amassed a total of 21.6 million views. Diet and alternative therapies were most common (15/30, 50%), which included recommendations to reduce hydrogenated oils, increase fruit and vegetable intake, and the use of traditional remedies such as garlic and black seed. Only 6.6% (2/30) of videos cited scientific literature. General cancer (15/30, 53%), breast (5/30, 17%), and cervical (4/30, 13%) cancers were most frequently mentioned. Doctors led 30% (9/30) of videos and were more likely to produce higher quality content, including significantly higher global quality scores (GQS=4, median 4, IQR 4-4 vs 3, median 3, IQR 2-3, P=.06). Over half of the videos had low understandability (16/30, 53%) and actionability (18/30, 60%). Emotionally framed content had the highest engagement across likes and shares, although this did not reach statistical significance (P=.08 and P=.05, respectively). However, emotional tone was significantly associated with lower GQS scores (P=.01). GPT-4 showed high agreement with human coders for cancer type (Cohen κ=1.0), strong agreement for GQS (κ=0.94), but low agreement for tone classification (κ=0.15), due to misclassification of emotional delivery from text-only input.

Conclusions: Arabic-language TikTok cancer prevention content is highly engaging but variable in quality, with emotionally framed videos attracting substantial attention despite lower informational value. Artificial intelligence-assisted tools show strong

potential for scalable, multilingual health content analysis, but multimodal approaches are needed to accurately interpret tonal and audiovisual features.

(JMIR Infodemiology 2026;6:e77888) doi:10.2196/77888

KEYWORDS

AI-driven content analysis; cancer prevention; engagement; digital health communication; TikTok

Introduction

Global Burden of Cancer

Worldwide, the cancer burden continues to rise, with an estimated 20 million new cases and 9.7 million deaths reported in 2022, projected to reach 35 million cases by 2050 [1]. The Arab world, which includes 22 countries, faces a particularly rapid increase in cancer incidence, with rates expected to rise 1.8-fold by 2030 [2]. Cancer ranks as a leading cause of death in many Arab nations, with Lebanon reporting the highest incidence of bladder cancer worldwide and Egypt contributing significantly to global liver cancer mortality [3,4]. Despite these rising trends, cancer prevention awareness remains limited, with low participation in screening programs and persistent misconceptions about cancer causes and treatment [5-8].

TikTok as a Source of Health Information

Social media platforms, such as TikTok, an emerging short-video app, have become major sources of health information worldwide [9]. During the COVID-19 pandemic, the platform saw a surge in health professionals and organizations using it to share medical knowledge and public health messages [10]. This shift highlighted the growing need for health care professionals to integrate video-based social media platforms, such as TikTok, into digital health communication strategies [11]. However, TikTok's global reach comes with region-specific challenges. Unlike other US-based platforms that apply universal moderation policies, TikTok uses localized moderation, tailoring its policies by region. This has raised concerns among Arabic-speaking users, particularly in North Africa, where dialect-specific moderation tools are lacking. Users often resort to strategies such as "algospeak" to avoid perceived censorship, and content moderation algorithms developed with limited dialect training data and nonnative annotators may misclassify or fail to flag harmful health misinformation. These dynamics highlight the urgent need to ensure equitable, culturally sensitive content governance as platforms such as TikTok become central to health communication ecosystems [12].

Arabic-Speaking Populations, an Understudied Demographic

Arabic-speaking populations, both in Arab countries and in diaspora communities, represent an understudied demographic in cancer prevention research [13,14]. Cultural beliefs, religious considerations, and misinformation often influence health behaviors, contributing to lower participation in preventive measures [15,16]. Language barriers further restrict access to reliable health information, with the Arabic language notably underrepresented in digital health research, along with the scarcity of validated Arabic-language health literacy tools and medical datasets [17-19]. This lack of research makes it challenging to assess the accuracy and effectiveness of Arabic-language health content, particularly on social media [20-22]. Understanding how cancer prevention messages are framed, their alignment with evidence-based guidelines, and their audience engagement is crucial for improving digital health communication among Arab nations and diaspora populations [23,24].

Large Language Models for Analyzing Digital Health Communication

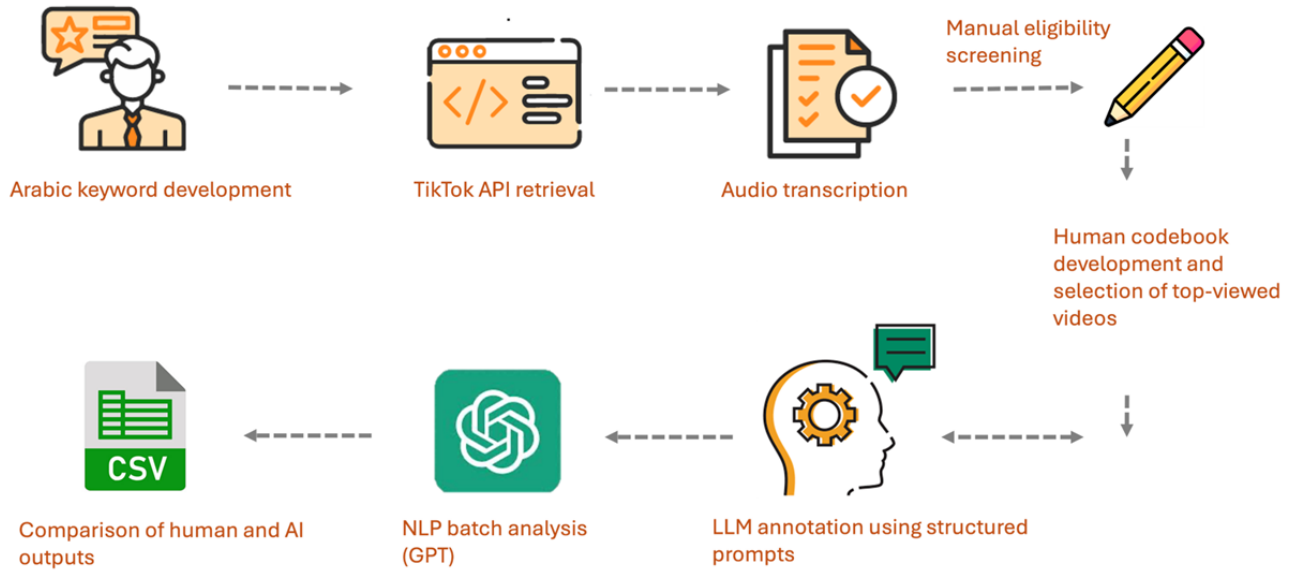
Advancements in artificial intelligence (AI), particularly large language models (LLMs) such as GPT, offer new opportunities for analyzing digital health communication in understudied languages [25,26]. GPT-4 has demonstrated high accuracy in detecting sentiment, misinformation, and medical accuracy across multiple languages [26]. Unlike traditional natural language processing tools, GPT does not require extensive language-specific training, making it a scalable tool for content analysis [26-28]. This study sought to examine TikTok videos on cancer prevention in Arabic, assess the content quality of the videos, and explore the role of LLMs such as GPT-4 in evaluating digital health content. By identifying gaps in digital health communication, this research seeks to inform strategies for improving cancer prevention awareness among Arabic-speaking communities.

Methods

Data Source and Retrieval

We used a multistep analytic workflow to identify, process, and analyze Arabic-language TikTok videos related to cancer prevention, integrating human coding with AI-assisted annotation. Figure 1 provides an overview of the full workflow, including keyword development, video retrieval, transcription, eligibility screening, manual coding, AI-based annotation, and assessment of agreement between human and AI outputs.

Figure 1. Overview of the analytic workflow: Arabic-language TikTok videos were retrieved using an iterative keyword strategy, transcribed, and screened for eligibility. A subset was used to develop the coding framework, after which the top 25% (N=30) most-viewed videos were manually coded and annotated using a large language model. Agreement between human and AI classifications was assessed using Cohen κ coefficient. Full methodological details are provided in the Methods section. AI: artificial intelligence; API: application programming interface; LLM: large language model; NLP: natural language processing.



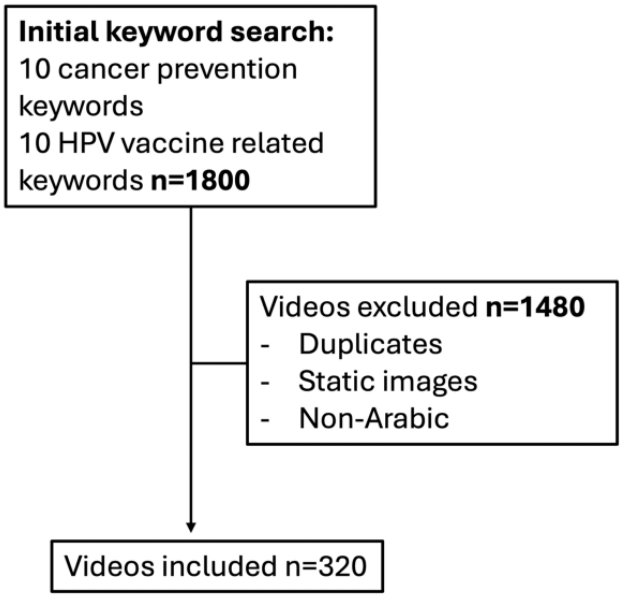
Using the TikTok application programming interface (API), we retrieved Arabic-language TikTok videos related to cancer prevention and the HPV vaccine from 2021 to 2024. This time frame was selected based on data from the Arab Youth Survey, which indicated an increasing trend in TikTok market penetration among young Arabs aged 18 to 24 years during this period. Specifically, daily TikTok usage more than doubled from 21% in 2020 to 50% in 2022, highlighting the platform's growing influence during this period [29]. Given that younger generations often play a key role in disseminating health information within their families, this period was considered optimal for capturing relevant content.

Search Strategy and Transcription

The research team developed an Arabic keyword (20 keywords) list focused on cancer prevention and HPV vaccination (Figure 1), which was iteratively refined and expanded using GPT-4 to ensure broad topical coverage (eg, "cancer prevention," "HPV vaccine," "tumor prevention," "healthy nutrition to prevent cancer," and "vaccination to protect against cervical cancer"). We then collected 1800 videos as above, excluding captions and comments (Figure 2). Videos were retrieved over several days because we were limited to 10,000 requests every 24 hours. Videos were transcribed using Sonix AI, selecting the Arabic transcription option. This Arabic transcription service is designed to handle a wide range of accents and regional dialects using advanced speech-to-text technology, though the model cannot be fine-tuned by users. All transcripts were manually reviewed for accuracy by the first author. No videos were truncated or cut during transcription, and the full audio of each TikTok was captured by Sonix's Arabic transcription. Sonix transcriptions focused on the lexical content of speech, the words and their semantic meaning, rather than acoustic properties such as pitch, pauses, or emphasis. As a result, nonverbal or visual information was not represented.

As part of our initial data collection using the TikTok API, we also retrieved publicly available aggregate engagement metrics for each video by country. These data, though not directly linked to the final analytic sample, provided insight into the geographic reach and broader engagement with Arabic-language cancer prevention content across global audiences. This included data on total views, likes, and shares by country, allowing us to assess which regions exhibited the highest levels of user interaction.

Figure 2. The filtering process used to identify top-performing Arabic-language TikTok videos related to cancer prevention. The final dataset included 30 videos in the top 25% most-viewed videos.



Codebook Development Using a Random 50-Video Subset

A random 16% (n=50) subset of the 320 videos was manually coded to create and refine the codebook. The research team reviewed coded data, discussed emerging categories, and reached consensus on definitions. The final codebook included classifications for content type, cancer type mentioned, tone and register, uploader type, presence of religious references, cautionary messages or disclaimers, target demographic, content quality indicators (global quality score and the Patient Education Materials Assessment Tool for Audiovisual Material [PEMAT AV]). Videos were coded as referencing scientific literature if they included direct citations or explicit mentions of scientific guidelines. This was used as a proxy for transparency, not as a definitive assessment of evidence-based accuracy. Tone and register were coded and simplified to 2 dimensions: emotional tone (eg, personal storytelling or emotional appeal, inclusive of expressive vocal delivery such as raised voice and dramatic pauses) and linguistic register (eg, casual or serious). This enabled consistency in both manual coding and AI-based classification and prioritized standardization over nuanced qualitative distinctions. If multiple tones were present within a video, the dominant style was recorded based on overall delivery. Videos were also coded for the specific types of cancer mentioned, including breast, colorectal, liver, pancreatic, brain, lung, cervical, oral, bladder, lymphoma, prostate, other cancers, and general cancer content (Textbox 1).

Textbox 1. Descriptions of content features and coding categories applied to Arabic TikTok videos.

Evidence-based

- Reference to a scientific study or cites scientific literature.

Emotional tone and linguistic register

- Describes how information is delivered:
  - Emotional tone: expresses personal experiences, urgency, or affective delivery. For example, “Dealing with this person causes cancer… it’s not food or drink that harms you; it’s people.”
  - Linguistic register: Casual: uses everyday, friendly language. For example, “I wish we could change this behavior, because it's literally a fountain of cancerous tumors.” Serious: uses formal or urgent phrasing. For example, “You have to come to the clinic. We have to do the mammogram. We have to take a sample. We have to..”

Content types

- Cancer prevention topics mentioned:
  - Diet alternative therapies: content promoting natural or nonclinical cancer prevention strategies, including the use of raw garlic, black seed, and dietary recommendations such as reducing hydrogenated oils.
  - Screening and early detection: mammograms and Papanicolaou test smears.
  - Vaccination: HPV vaccine.
  - Self-examination and symptom awareness: breast or testicular checks.
  - Smoking cessation: avoiding tobacco.
  - Stress or negativity: links between stress and cancer.
  - Survivor experience: sharing stories of cancer survival.
  - Chemical carcinogens: mention of environmental or food-based chemicals.

Cancers mentioned

- Specific and nonspecific cancer types cited, including:
  - General, breast, colorectal, liver, pancreatic, brain, lung, cervical, oral, bladder, lymphoma, prostate, other, and no cancer mentioned.

Speaker (doctor, self ID, or layperson)

- Whether the speaker self-identifies as a doctor (based on credentials in profile or linked accounts), self-identifies without affiliation, or does not claim any medical background.

Religious reference

- Mentions religious texts or spiritual framing of health advice.

Cautionary message or disclaimer

- Debunks a myth, adds a disclaimer, or highlights risks.

Target demographic

- Intended audience includes women, men, both genders, and youth.

Patient Education Materials Assessment Tool (PEMAT) understandability

- Percentage score based on clarity and ease of understanding.
- Coded as: high: 67%-100%, medium: 34%-66%, and low: 0%-33%.

PEMAT actionability

- Percentage score based on clear steps for action. Same categorization as understandability.

Global Quality Scale (GQS)

- 1=very poor (not useful) to 5=excellent (highly useful and comprehensive).

The coding framework included an assessment of uploader type which was classified into three categories: (1) doctors, whose credentials were corroborated through profile information (eg, "Dr" in username or bio); (2) self-identified doctors, who claimed a medical background without confirmable credentials; and (3) laypersons, with no stated or apparent medical affiliation. When the uploader status was ambiguous, the research team reviewed TikTok biographies, posted video content, and any linked social media profiles to determine affiliation. Clinic-affiliated accounts with a medical focus were also coded as doctor-led. Influencer status was assessed separately based on follower count, due to limited public information or unidentifiable handles. Creators with 100,000 or more followers were classified as influencers, regardless of medical background or professional identity. Videos were also coded for references to religious texts or beliefs, including phrases that framed health outcomes as divinely guided (*qadr*). Cautionary messages were defined as explicit statements that debunk myths, provide disclaimers, or warn against specific risks, such as ensuring that individuals with specific conditions avoid potential harms associated with alternative therapies or clarifying that the HPV vaccine is not exclusively for girls.

The target demographic of each TikTok video was categorized based on direct mention of the audience in the video, such as if the creator explicitly addressed a specific group (women or men), the reference to a specific age group (young people), or broad health advice, which was categorized as both genders.

Sample Selection Based on the 75th Percentile Cutoff

To focus on the most visible content, we selected the top 25% (N=30) most-viewed videos for detailed manual and AI analysis. To identify the most engaged content, we applied a 75th percentile cutoff. The decision to use the 75th percentile was based on established statistical principles for performance classification, a method that aligns with industry standards and prior studies that have used percentile-based cutoffs (75th percentile) to differentiate high-performing content from lower-engagement material [30]. This method helps mitigate the influence of outliers while allowing for the analysis of content that drives interaction and user engagement. By applying this approach, we ensured that this subset of videos reflected the most influential cancer prevention messaging on TikTok, aligning with established research methodologies in social media health communication [30].

Manual Coding and Interreliability Testing

The coding of the top 25% (N=30) of most-viewed videos was independently conducted by 2 study team members (AK and LS). The 2 coders met to finalize the codebook and resolve any coding discrepancies. To ensure coding reliability, interrater agreement was assessed using Cohen κ. Cohen κ was calculated for each coding category before reconciliation. Cohen κ values were interpreted using the following standard: values below 0.20 indicated slight agreement; 0.21-0.40, fair agreement; 0.41-0.60, moderate agreement; 0.61-0.80, substantial agreement; and 0.81-1.00, almost perfect agreement.

Assessment of Understandability, Actionability, and Quality

The PEMAT AV was used to assess the understandability and actionability of the videos [31]. The understandability section contains 13 items, and the actionability section includes 4 items, which can each be scored as 0 ("disagree"), 1 ("agree"), or "not applicable." For each section, PEMAT AV scores are calculated as percentages by dividing the points achieved by the items evaluated for the video. Therefore, higher values are indicative of higher understandability and actionability. The PEMAT AV has been widely used to evaluate health communication materials across formats, including videos, animations, and patient education modules for a range of topics, including chronic disease management, vaccine education, cancer prevention, and health literacy interventions [32,33]. We dichotomized PEMAT AV scores with 0%-66% considered as "low understandability," and 67%-100% as "high understandability." This threshold was based on the original guidance provided by Shoemaker et al [31], who recommended 70% as a benchmark for acceptable educational materials. The Global Quality Scale (GQS) was used to evaluate the overall quality, flow, and usefulness of each video's health information. This 5-point Likert scale has been widely used to evaluate the reliability and educational value of online medical and public health content [34]. The score represents the perception of the trained coder (in our case, 2 Arabic-speaking coders with experience evaluating health communication content). The GQS is scored based on the following scale 1="very poor quality, missing information, not useful"; 2="generally poor quality, some missing information, very limited use"; 3="moderate quality, some information adequately discussed, somewhat useful"; 4="good quality, most relevant information discussed, useful"; and 5="excellent quality, all relevant information discussed, very useful" [35].

AI-Based Annotation

To generate AI-based annotations, we used a 1-shot prompting approach, in which a single, structured prompt was provided to the model to classify each video based on predefined categories from our manually generated codebook [36]. The prompt included clear definitions to guide the model's interpretation. For instance, the model was instructed as follows: "Answer the questions as precisely and faithfully as possible using the provided context. The provided text is in Arabic with various dialects. Provide the answers in JSON format. Ensure that all responses are directly based on the provided text without assumptions or external information."

Questions included items such as "List any cancers mentioned in the text: Options: General, Breast, Colorectal, Liver, Pancreatic, Brain, Lung, Cervical, Oral, Bladder, Lymphoma, Prostate, Other and No cancer mentioned." For GQS scoring: "A Global Quality Score (GQS): Options: a score from 1 to 5 with:1: Poor quality, poor flow, and not useful 2: Generally poor quality and flow, but some information is listed 3: Moderate quality and flow, but some important information is poorly discussed, 4: Good quality and flow, but some topics are not covered 5: Excellent quality and flow, and very useful."

For each coding category, such as cancer type or GQS score, the model was instructed to return 1 label per video, such as output the answers in the following JSON format: “cancers_mentioned”: [<list from cancers list>], “GQS_Score”: “<one score>”

AI outputs were generated in Python (Python Software Foundation) through a batch analysis pipeline [37]. The GPT model was optimized using iterative prompt engineering, refining the input structure to improve consistency in classification and fidelity to the codebook. This enabled efficient, scalable annotation of video characteristics while minimizing ambiguity, and facilitated direct comparison between AI-generated and human-coded classifications. AI-generated outputs were reviewed by human coders and systematically compared to manual annotations for the top-viewed videos. Cohen κ was used to evaluate interrater reliability between human and AI classifications. This analysis was conducted using the Cohen kappa score function in Python, which measures agreement beyond chance for categorical variables.

Statistical Analysis

As this was an exploratory analysis, GPT-generated annotations were limited to 3 key categories: cancer type, tone and register, and GQS score. Descriptive statistics were used to summarize video characteristics, and inferential analyses were conducted using nonparametric tests due to high variability in the data. The Wilcoxon Rank Sum test was used to compare median engagement metrics (likes and shares) across groups, as engagement data were highly skewed. Fisher Exact Tests were used for categorical comparisons where sample sizes were small or expected cell counts were low. These statistical methods were selected to ensure robustness despite nonnormal distributions and heterogeneous group sizes.

Ethical Considerations

This study analyzed publicly available TikTok videos related to cancer prevention using the TikTok Research API and did not involve direct interaction with human participants. No private, identifiable, or nonpublic user data were collected. All data were accessed and analyzed in accordance with TikTok’s terms of service and research data use policies.

Results

Overview

The final analytic dataset included 30 TikTok videos, representing the top 25% most-viewed content from an initial pool of 320 Arabic-language videos related to cancer prevention (cutoff at 59,640 views). These 30 videos collectively amassed 21.6 million views, 445,000 likes, and 146,000 shares (see Tables 1 and 2).

Table 1. Characteristics of TikTok videos (N=30).

Characteristic	Values, n (%)
Content type	
Diet and alternative therapies	15 (50)
Screening and early detection	6 (20)
HPV[a] vaccination	3 (10)
Self-examination and symptoms to look out for	1 (3)
Smoking cessation	1 (3)
Stress and negativity	1 (3)
Survivor experience	1 (3)
Chemical carcinogens	1 (3)
Cancers mentioned	
General cancer	15 (50)
Breast cancer	6 (20)
Cervical cancer	4 (13)
Colon cancer	2 (7)
Bladder cancer	1 (3)
Multiple cancers	1 (3)
Testicular cancer	1 (3)
Emotional tone and register	
Casual	16 (53)
Serious	8 (27)
Emotional	6 (20)
Target demographic	
Both genders	19 (63)
Women	8 (27)
Men	1 (3)
Young people	2 (7)
Led by doctors (corroborated or self-identified)	
Yes (credentials corroborated)	9 (30)
Yes (self-identified or no confirmable credentials)	6 (20)
No (layperson did not state medical affiliation)	13 (43)
Medical clinic affiliated	2 (7)
Evidence-based	
Yes	2 (7)
No	28 (93)
Cautionary message or disclaimer	
No	16 (53)
Yes	14 (47)
PEMAT[b] understandability	
High (≥67%)	14 (47)
Low (≤66%)	16 (53)
PEMAT actionability	

Characteristic	Values, n (%)
High (≥67%)	15 (50)
Low (≤66%)	15 (50)
Religious reference	
Yes	6 (20)
No	24 (80)
GQS[c]	
1 (very poor)	1 (3)
2 (poor)	5 (17)
3 (moderate)	7 (20)
4 (good)	17 (60)
5 (excellent)	0 (0)

[a]HPV: human papillomavirus.

[b]PEMAT: Patient Education Materials Assessment Tool.

[c]GQS: Global Quality Scale.

Table 2. Other characteristics of TikTok videos (N=30).

Engagement	Minimum-maximum	Median (IQR)
Like count	524-116,493	3062 (1370-18,629)
Share count	37-36,403	751.5 (199-4019)
View count	59,116-8,490,149	176,391 (92,166-592,176)

Emotional Tone and Linguistic Register

Casual was the most common (16/30, 53%) code, followed by serious (8/30, 27%) and emotional (6/30, 20%). Emotional videos were more engaging than others, receiving the highest median likes and share counts. However, emotional videos were associated with lower global quality scores (median 2, IQR 2-3), while serious and casual videos received higher scores (median 4, IQR 4-4). The difference in GQS across tones was statistically significant (P=.01), indicating that higher engagement did not correspond with higher content quality.

Content Types

The most common content was diet and alternative therapies (15/30, 50%), including content promoting the use of raw garlic, black seeds, or reducing hydrogenated oils. This was followed by screening and early detection (6/30, 20%) and HPV vaccination (3/30, 10%). Other content types, including self-examination, stress, smoking cessation, and survivor stories, each appeared in only 1 of 30 (3%) videos. The videos in the top 25% (N=30) most-viewed videos focused on HPV vaccination, and those that mentioned cervical cancer were created by self-identified doctors, including 2 verified and 1 unverified account. Two used a serious tone, and 1 used a casual tone. All 3 videos received a global quality score (GQS) of 4.

Cancers Mentioned

The most frequently mentioned cancers were general cancer (16/30, 53%), with breast cancer (5/30, 17%) and cervical cancer (4/30, 13%) commonly referenced. Less frequently mentioned were colon (2/30, 7%), bladder (1/30, 3%), multiple cancers (1/30, 3%), and testicular cancer (1/30, 3%).

Speaker (Doctor, Self-Identified, or Layperson)

Among the top 25% (N=30) most-viewed videos, 17 (57%) were led by individuals identifying as doctors, including 9 (30%) verified doctors, 6 (20%) unverified, and 2 (7%) accounts affiliated with medical clinics. The remaining 13 of 30 (43%) accounts were led by laypeople. Of the 17 doctor-led accounts, 8 (47%) verified and 2 (12%) unverified accounts met the threshold for influencer status. Among the 13 laypeople accounts, 4 (31%) accounts met the influencer criteria. While only 4 of the 13 nondoctor-led accounts met the influencer threshold (≥100,000 followers), 3 additional laypeople's accounts had a substantial following between 20,000 and 60,000 followers.

Target Demographics, Religious Reference, and Presence of Cautionary Message

Religious framing appeared in 6 of 30 (20%) videos, with references to divine (*qadr*) will or spiritual health advice. Further, 14 of the 30 (47%) videos included a cautionary message or disclaimer, such as warnings about misinformation or clarification on cancer risk factors. Most videos targeted both genders (19/30, 63%), followed by women (8/30, 27%), young people (2/30, 7%), and men (1/30, 3%).

Evidence-Based, Patient Education Materials Assessment Tool: Understandability and Actionability

Only 2 (7%) of the top 25% (N=30) most-viewed videos explicitly cited scientific literature or guidelines. A total of 53%

(16/30) of videos were rated low on understandability (score ≤66%). Notably, all 6 doctor-led videos promoting diet and alternative therapies scored high (≥67%) for understandability, while all 9 layperson videos in that category scored low. Furthermore, 50% (15/30) of videos were rated low on actionability. Among diet and alternative health-related videos, those led by doctors were more likely to be actionable by 83% (25/30) compared to those led by laypeople (17/30, 56.6%), though this difference was not statistically significant (P=.58).

About GQS

A total of 60% (18/30) of videos were rated as good (score of 4), 23% (7/30) as moderate (score of 3), and 17% (5/30) as poor (score of 2). Only 1 (3%) video was rated very poor (score of 1), and none were rated excellent. Videos led by doctors promoting diet and alternative therapies had significantly higher GQS scores than those led by laypeople (P=.06).

Human and AI Agreement

Agreement between human coders was high across all domains (κ=0.84). There was perfect agreement between human and AI annotations for cancer type (κ=1.0) and strong agreement for GQS scoring (κ=0.94), though most discrepancies occurred between scores of 3 (moderate) and 4 (good), indicating difficulty distinguishing between mid and high-quality content. Agreement was lower for tone classification (κ=0.15), with AI misclassifying emotional delivery when relying on text-based input alone.

Geographic Reach of Arabic-Language Cancer Prevention Content on TikTok

TikTok platform data indicated that Arabic-language cancer prevention content generated substantial engagement from users in both Arab-majority countries (eg, Egypt, Jordan, and Saudi Arabia) and diaspora contexts such as the United States, France, and Germany. The United States ranked in the top 10 for total views, highlighting the global reach of Arabic-language cancer messaging.

There was high agreement in cancer type between human and AI annotations (κ=1.0), and similarly high agreement in GQS scoring (κ=0.94). Tone classification showed lower concordance. While the AI model correctly identified many casual and serious tones, it misclassified emotional content in several cases, resulting in slight overall agreement (κ=0.15). Manual review confirmed that GPT-based annotation performed reliably across dialectal variations from multiple Arabic-speaking countries, indicating that cross-dialect consistency is achievable when coupled with human verification.

Discussion

Principal Findings

This study makes 3 contributions to the broader literature on online health videos and TikTok specifically. First, it provides the first systematic analysis of Arabic-language TikTok videos on cancer prevention. Second, it identifies an engagement quality gap in Arabic language cancer prevention content, extending prior English-language findings that emotionally charged posts often receive higher engagement [38]. Third, the present study advances methodological research by evaluating GPT-4's performance on Arabic transcript-only inputs from short-form videos: the model demonstrated high reliability for structured categorical variables but low reliability for tone classification, underscoring the need for multimodal approaches that incorporate audio and visual cues. Together, these contributions deepen the understanding of Arabic-language TikTok health communication and illustrate both the potential and current limitations of AI-assisted content analysis across global digital ecosystems. While prior work has shown that LLMs can support qualitative researchers by generating themes from social media corpora in a single prompt [39], their use for systematic content analysis of short-form video data has not, to our knowledge, been previously demonstrated.

Within our sample, videos promoting diet and alternative therapies were among the most viewed. Studies in other cultural contexts have similarly shown that traditional or community-based health guidance often thrives because it is relational, linguistically resonant, and perceived as more trustworthy than institutional messages [40]. Research on TikTok in English more broadly echoes this pattern: a content analysis of health-related "EduTok" videos found that audiences most frequently engaged with educational posts related to diet, exercise, and sexual health, suggesting consistent user interest in familiar, lifestyle-oriented themes [41]. Together, these parallels suggest that what circulates widely on Arabic TikTok may reflect a broader sociocultural logic in which familiarity and affective connection drive credibility and engagement. It is important to note, however, that patterns related to content type and cancer type in this study are driven primarily by a small number of highly represented categories (eg, diet and alternative therapies and screening and early detection), which reflects the distribution of high-engagement Arabic-language TikTok content rather than a comprehensive representation of cancer prevention topics.

Beyond these general patterns, our data illustrate how an engagement quality gap appears specifically within Arabic TikTok cancer prevention content. Emotional tone was associated with higher engagement, even when informational quality was low. One widely circulated video, for example, claimed that "toxic people, not food or genetics," cause cancer, an emotionally resonant but scientifically inaccurate claim that drew considerable engagement. These findings align with prior TikTok-specific studies showing that affectively charged content outperforms factual or instructional posts [41]. Importantly, our results do not imply that emotional tone alone determines virality; rather, they suggest that emotional framing, creator identity, and algorithmic amplification together create conditions in which lower-quality but more affectively engaging messages can spread widely.

Targeting of young people was limited despite TikTok's prominence among youth. Only a small proportion of high-engagement videos explicitly addressed adolescents or young adults, even though early-life behaviors, such as HPV vaccination, tobacco use, diet, and physical activity, are critical for cancer prevention. The absence of youth-directed content suggests a missed opportunity to leverage TikTok as a public health tool for early prevention messaging. Instead, content

often targeted adult women or general audiences, which may reflect creator demographics or cultural communication norms. However, it is important to recognize that reliance on TikTok for health information is not limited to adolescents. Many Arabic speakers in diaspora contexts, including Arab Americans, turn to social media due to linguistic and cultural barriers in traditional health care settings [17]. Consistent with this, our platform data showed high engagement from diaspora countries, including the United States, emphasizing TikTok's role as a transnational source of Arabic language health information. Furthermore, prior research has shown that immigrants frequently rely on online platforms for relatable and accessible health content, making the quality of digital communication a critical equity issue [42]. These patterns parallel findings from US studies showing that African American and Hispanic adults were more likely than White adults to seek health information through social media during the COVID-19 pandemic, underscoring how communication inequities can drive platform reliance among marginalized groups [42]. Future research should examine whether these same patterns extend to Arabic-language health content on other short-form video platforms such as Instagram Reels (Meta), YouTube Shorts (Google LLC), Facebook Watch (Meta), and Snapchat Spotlight (Snap Inc), which share similar algorithmic dynamics but may differ in moderation and audience reach.

Most of the analyzed videos lacked references to peer-reviewed literature or established clinical guidelines, and only 30% (9/30) were led by doctors (whose credentials could be corroborated). It is important to distinguish between being evidence-based and citing sources. While a video may communicate content that aligns with scientific consensus, the absence of explicit references may reduce credibility, especially in digital environments where users rely on transparency to assess trustworthiness. Although doctor-led videos produced higher quality content on average (eg, higher GQS scores), professional identity alone did not ensure high understandability or actionability (high understandability suggests that most viewers, including those with limited health literacy, can grasp the essential messages being communicated). This is particularly salient given that populations with lower health literacy are more likely to rely on TikTok for health information [43].

Interpreting PEMAT AV and GQS together provides important insight into the quality of doctor-led content. PEMAT AV, which is validated for audiovisual materials, assesses whether information is communicated clearly and whether viewers are given actionable guidance, whereas GQS reflects a broader, more subjective appraisal of overall informational quality and usefulness. These differences are therefore meaningful rather than contradictory and underscore the value of using PEMAT AV and GQS as complementary measures when evaluating short-form health content [33]. Future studies may benefit from incorporating additional quality frameworks to further capture dimensions of informational rigor and communicative nuance.

Our findings show that doctor-led videos achieved higher quality scores but did not generate comparable engagement. This aligns with prior research on Arabic language health content across other platforms. For instance, studies of Arabic breast cancer videos on YouTube have shown that videos produced by trusted institutions tend to be more accurate but far less popular than those by individual users. This recurring pattern across platforms suggests that credibility alone does not guarantee visibility, a consistent challenge in health communication on social media. Effective health communication may therefore require pairing evidence-based content with narrative appeal, cultural resonance, and accessible delivery formats to compete with misinformation and emotionally engaging but lower-quality material.

Content moderation practices also play a role in shaping what health information circulates across Arabic-speaking regions. Unlike platforms that apply uniform global policies, TikTok relies on region-specific moderation teams and language filters, which may inconsistently flag or downrank health misinformation. While this study focused on cancer prevention, the engagement and credibility patterns we observed echo those reported in Arabic language TikTok vaccine content, suggesting that visibility dynamics may be shaped more by platform design and algorithmic incentives than by the specific health topic itself [22].

Finally, our study underscores the promise of LLMs for scalable analysis of health-related content in underrepresented languages. GPT-based coding achieved high reliability in classifying categorical variables such as cancer type and video quality (κ=0.94 to κ=1.0). However, it performed less consistently in detecting tone and register, particularly emotional delivery.

This finding stands in partial contrast to prior work, which has reported strong LLM performance in multilingual sentiment analysis of social media content [26]. A key difference, however, lies in the approach, which used large-scale labeled data from high-resource, text-based platforms. In contrast, our 1-shot method relied on transcripts of Arabic TikTok videos, where much of the emotional tone is conveyed through audiovisual cues such as intonation and facial expression not captured in text alone.

Emerging multimodal AI systems such as Gemini (Google LLC) and Google Cloud Video Intelligence, which can jointly interpret audio, visual, and textual inputs, hold promise for overcoming these limitations and enabling more contextually accurate annotation of short-form health content across languages. Applied tools such as ScreenApp, which automatically integrates speech-to-text, speaker detection, and scene-level video analysis, further illustrate how multimodal pipelines are already being used to extract meaningful patterns from audiovisual content [44].

The observed pattern in our findings points to 2 directions for future work: (1) developing a labeled Arabic TikTok dataset to refine tone detection in LLMs, and (2) adopting multimodal pipelines that combine audio, visual, and text cues to better capture emotion and on-screen gestures. This hybrid approach, using LLMs for large-scale triage and multimodal or human review for nuanced interpretation, offers a scalable path for improving health content analysis across languages.

These findings have important implications for public health outreach in Arabic-speaking communities, both within the Arab world and across diaspora populations. Given the limited availability of culturally tailored, Arabic-language health

education materials and the growing reliance on social media for information, TikTok represents both a powerful tool and a potential vector for misinformation. While it can amplify accurate messaging, it also enables the rapid spread of misinformation. Addressing this will require multipronged strategies: empowering health care providers with the skills to create engaging content, leveraging AI for content monitoring, and partnering with trusted community figures to amplify reliable messages. Future interventions may leverage narrative and emotionally resonant formats to pair evidence-based content with styles that match user engagement preferences.

Limitations

This study has several limitations. First, it focused on videos in the top 25% (N=30) most-viewed, which introduces an engagement bias and thus may not fully represent the broader landscape of Arabic-language cancer prevention content on TikTok. In addition, AI comparison was applied only to these top-viewed videos, rather than the full set of 320 eligible videos, limiting our ability to fully assess the scalability and generalizability of AI-based annotation across the entire dataset.

Second, we did not analyze or control for video length, which may influence engagement metrics. However, because the sample was drawn from the top 25% (N=30) of most-viewed videos, it likely reflects videos optimized for typical TikTok viewing behavior, minimizing major variability in duration effects. Third, while the Patient Education Materials Assessment Tool and GQS frameworks are validated tools, they may not fully capture the stylistic and communicative nuances characteristic of short-form, audiovisual social media content. Fourth, to facilitate AI-based classification, tone and linguistic register were simplified into broad categories, which may have limited the detection of the more subtle or culturally embedded communication styles typically captured through qualitative analysis. Fifth, although use of the TikTok API helped reduce algorithmic sampling bias, platform-specific ranking mechanisms and personalization features may still have influenced which videos achieved high visibility. In addition, analyses of content type and cancer type were concentrated within a narrow subset of categories due to the limited representation of other topics, and findings should therefore not be generalized to less frequently represented cancers or prevention behaviors. Finally, the AI model operated solely on transcribed audio, analyzing text without access to visual or prosodic cues such as facial expressions, gestures, or intonation, elements that are often common in video-based communication.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Conflicts of Interest

None declared.

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024;74(3):229-263 [FREE Full text] [doi: 10.3322/caac.21834] [Medline: 38572751]
2. Al-Muftah M, Al-Ejeh F. Cancer incidence and mortality estimates in Arab countries in 2018: a GLOBOCAN data analysis. Cancer Epidemiol Biomarkers Prev 2023;32(12):1738-1746 [FREE Full text] [doi: 10.1158/1055-9965.EPI-23-0520] [Medline: 37733340]
3. Arafa MA, Rabah DM, Farhat KH. Rising cancer rates in the Arab world: now is the time for action. East Mediterr Health J 2020;26(6):638-640 [FREE Full text] [doi: 10.26719/emhj.20.073] [Medline: 32621496]
4. Abbas NF, Aoude MR, Kourie HR, Al-Shamsi HO. Uncovering the epidemiology of bladder cancer in the Arab world: a review of risk factors, molecular mechanisms, and clinical features. Asian J Urol 2024 Jul;11(3):406-422 [FREE Full text] [doi: 10.1016/j.ajur.2023.10.001] [Medline: 39139531]
5. Sayan M, Eren AA, Tuac Y, Langoe A, Alali B, Aynaci O, et al. Prostate cancer awareness in the middle east: a cross-sectional international study. JCO Glob Oncol 2024;10:e2400171. [doi: 10.1200/GO.24.00171] [Medline: 38991182]
6. Ahmed HAA, Abbas MH, Hussein HA, Nasr RSF, Lashen AA, Khaled H, et al. Correction: cervical cancer screening uptake in Arab countries: a systematic review with meta-analysis. BMC Cancer 2024;24(1):1478 [FREE Full text] [doi: 10.1186/s12885-024-13233-2] [Medline: 39614258]
7. Marouf A, Tayeb R, Alshehri GD, Fatani HZ, Nassif MO, Farsi AH, et al. Public perception of common cancer misconceptions: a nationwide cross-sectional survey and analysis of over 3500 participants in Saudi Arabia. J Family Med Prim Care 2023;12(6):1125-1132. [doi: 10.4103/jfmpc.jfmpc_1753_22] [Medline: 37636192]
8. Elshami M, Naji SA, Dwikat MF, Al-Slaibi I, Alser M, Ayyad M, et al. Myths and common misbeliefs about colorectal cancer causation in Palestine: a national cross-sectional study. JCO Glob Oncol 2024;10:e2300295 [FREE Full text] [doi: 10.1200/GO.23.00295] [Medline: 38166235]
9. Kirkpatrick CE, Lawrie LL. TikTok as a source of health information and misinformation for young women in the United States: survey study. JMIR Infodemiology 2024;4:e54663 [FREE Full text] [doi: 10.2196/54663] [Medline: 38772020]

10. Basch CH, Hillyer GC, Jaime C. COVID-19 on TikTok: harnessing an emerging social media platform to convey important public health messages. Int J Adolesc Med Health 2022;34(5):367-369 [FREE Full text] [doi: 10.1515/ijamh-2020-0111] [Medline: 32776899]
11. Song S, Xue X, Zhao YC, Li J, Zhu Q, Zhao M. Short-video apps as a health information source for chronic obstructive pulmonary disease: information quality assessment of TikTok videos. J Med Internet Res 2021;23(12):e28318 [FREE Full text] [doi: 10.2196/28318] [Medline: 34931996]
12. Elswah M. Moderating Maghrebi Arabic Content on Social Media. URL: https://cdt.org/wp-content/uploads/2024/09/2024-09-26-CDT-Research-Global-South-Moderating-Report-English-Arabic-final.pdf [accessed 2026-01-20]
13. Zaidi M, Fantasia HC, Ahmed R, Lee DN, Valdman O, Poghosyan H, et al. Experiences with cancer screenings among arabic-speaking refugee women. Nurs Womens Health 2025;29(2):109-119. [doi: 10.1016/j.nwh.2024.09.004] [Medline: 39947246]
14. Bhattacharya R, Chen N, Shim I, Kuwahara H, Gao X, Alkuraya FS, et al. Massive underrepresentation of Arabs in genomic studies of common disease. Genome Med 2023;15(1):99 [FREE Full text] [doi: 10.1186/s13073-023-01254-8] [Medline: 37993966]
15. Hammoud MM, White CB, Fetters MD. Opening cultural doors: providing culturally sensitive healthcare to Arab american and American muslim patients. Am J Obstet Gynecol 2005;193(4):1307-1311. [doi: 10.1016/j.ajog.2005.06.065] [Medline: 16202719]
16. Abboud S, De Penning E, Brawner BM, Menon U, Glanz K, Sommers MS. Cervical cancer screening among Arab women in the United States: an integrative review. Oncol Nurs Forum 2017;44(1):E20-E33 [FREE Full text] [doi: 10.1188/17.ONF.E20-E33] [Medline: 27991600]
17. Al-Jumaili AA, Ahmed KK, Koch D. Barriers to healthcare access for Arabic-speaking population in an English-speaking country. Pharm Pract (Granada) 2020;18(2):1809 [FREE Full text] [doi: 10.18549/PharmPract.2020.2.1809] [Medline: 32477432]
18. Bergman L, Nilsson U, Dahlberg K, Jaensson M, Wångdahl J. Health literacy and e-health literacy among Arabic-speaking migrants in Sweden: a cross-sectional study. BMC Public Health 2021;21(1):2165 [FREE Full text] [doi: 10.1186/s12889-021-12187-5] [Medline: 34823499]
19. Reich H, Hegerl U, Rosenthal A, Allenhof C. Arabic-language digital interventions for depression in German routine health care are acceptable, but intervention adoption remains a challenge. Sci Rep 2024;14(1):12097 [FREE Full text] [doi: 10.1038/s41598-024-62196-8] [Medline: 38866810]
20. Alfaris E, Alhazzani Y, Alkhenizan A, Irfan F, Almoneef N, Alyousefi N, et al. Assessing the validity of health messages used by the Saudi public in WhatsApp. Patient Prefer Adherence 2023;17:67-73 [FREE Full text] [doi: 10.2147/PPA.S397661] [Medline: 36632071]
21. Alammari OB, Odeh O, Alotaibi R, Zamzam M, Alghamdi AS. Assessment of the most popular educational content in Arabic available through four social media applications in Saudi Arabia that explains idiopathic clubfoot deformity. Cureus 2025;17(1):e76976. [doi: 10.7759/cureus.76976] [Medline: 39912023]
22. Sallam M, Al-Mahzoum K, Alkandari L, Shabakouh A, Shabakouh A, Ali A, et al. Descriptive analysis of TikTok content on vaccination in Arabic. AIMS Public Health 2025;12(1):137-161 [FREE Full text] [doi: 10.3934/publichealth.2025010] [Medline: 40248416]
23. Plackett R, Kaushal A, Kassianos AP, Cross A, Lewins D, Sheringham J, et al. Use of social media to promote cancer screening and early diagnosis: scoping review. J Med Internet Res 2020;22(11):e21582 [FREE Full text] [doi: 10.2196/21582] [Medline: 33164907]
24. Conley CC, Otto AK, McDonnell GA, Tercyak KP. Multiple approaches to enhancing cancer communication in the next decade: translating research into practice and policy. Transl Behav Med 2021;11(11):2018-2032 [FREE Full text] [doi: 10.1093/tbm/ibab089] [Medline: 34347872]
25. Lu Z, Peng Y, Cohen T, Ghassemi M, Weng C, Tian S. Large language models in biomedicine and health: current research landscape and future directions. J Am Med Inform Assoc 2024;31(9):1801-1811. [doi: 10.1093/jamia/ocae202] [Medline: 39169867]
26. Rathje S, Mirea D, Sucholutsky I, Marjieh R, Robertson CE, Van Bavel JJ. GPT is an effective tool for multilingual psychological text analysis. Proc Natl Acad Sci U S A 2024;121(34):e2308950121 [FREE Full text] [doi: 10.1073/pnas.2308950121] [Medline: 39133853]
27. Miao H, Li C, Wang J. A future of smarter digital health empowered by generative pretrained transformer. J Med Internet Res 2023;25:e49963 [FREE Full text] [doi: 10.2196/49963] [Medline: 37751243]
28. Garbarino S, Bragazzi NL. Evaluating the effectiveness of artificial intelligence-based tools in detecting and understanding sleep health misinformation: comparative analysis using Google Bard and OpenAI ChatGPT-4. J Sleep Res 2024;33(6):e14210. [doi: 10.1111/jsr.14210] [Medline: 38577714]
29. Survey TAY. URL: https://arabyouthsurvey.com/en/ [accessed 2026-01-20]
30. Klassen KM, Borleis ES, Brennan L, Reid M, McCaffrey TA, Lim MS. What people "like": analysis of social media strategies used by food industry brands, lifestyle brands, and health promotion organizations on Facebook and Instagram. J Med Internet Res 2018;20(6):e10227 [FREE Full text] [doi: 10.2196/10227] [Medline: 29903694]

31. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014;96(3):395-403 [FREE Full text] [doi: 10.1016/j.pec.2014.05.027] [Medline: 24973195]
32. Furukawa E, Okuhara T, Liu M, Okada H, Kiuchi T. Evaluating online and offline health information with the patient education materials assessment tool: protocol for a systematic review. JMIR Res Protoc 2025;14:e63489 [FREE Full text] [doi: 10.2196/63489] [Medline: 39813665]
33. Dimitroyannis R, Fenton D, Cho S, Nordgren R, Pinto JM, Roxbury CR. A social media quality review of popular sinusitis videos on TikTok. Otolaryngol Head Neck Surg 2024;170(5):1456-1466. [doi: 10.1002/ohn.688] [Medline: 38431902]
34. Gudapati JD, Franco AJ, Tamang S, Mikhael A, Hadi MA, Roy V, et al. A study of global quality scale and reliability scores for chest pain: an Instagram-post analysis. Cureus 2023;15(9):e45629 [FREE Full text] [doi: 10.7759/cureus.45629] [Medline: 37868472]
35. Li M, Yan S, Yang D, Li B, Cui W. YouTube as a source of information on food poisoning. BMC Public Health 2019;19(1):952 [FREE Full text] [doi: 10.1186/s12889-019-7297-9] [Medline: 31311523]
36. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023;25:e50638 [FREE Full text] [doi: 10.2196/50638] [Medline: 37792434]
37. Atkinson CF. AI-pocalypse now: automating the systematic literature review with SPARK (Systematic Processing and Automated Review Kit) - gathering, organising, filtering, and scaffolding. MethodsX 2025;14:103129 [FREE Full text] [doi: 10.1016/j.mex.2024.103129] [Medline: 39846014]
38. Abbas MJ, Khalil LS, Haikal A, Dash ME, Dongmo G, Okoroha KR. Eliciting emotion and action increases social media engagement: an analysis of influential orthopaedic surgeons. Arthrosc Sports Med Rehabil 2021;3(5):e1301-e1308 [FREE Full text] [doi: 10.1016/j.asmr.2021.05.011] [Medline: 34712967]
39. Deiner MS, Honcharov V, Li J, Mackey TK, Porco TC, Sarkar U. Large language models can enable inductive thematic analysis of a social media corpus in a single prompt: human validation study. JMIR Infodemiology 2024;4:e59641 [FREE Full text] [doi: 10.2196/59641] [Medline: 39207842]
40. Cruz ML, Christie S, Allen E, Meza E, Nápoles AM, Mehta KM. Traditional healers as health care providers for the Latine community in the United States, a systematic review. Health Equity 2022;6(1):412-426 [FREE Full text] [doi: 10.1089/heq.2021.0099] [Medline: 35801152]
41. O'Donnell N, Jerin SI, Mu D. Using TikTok to educate, influence, or inspire? A content analysis of health-related EduTok videos. J Health Commun 2023;28(8):539-551. [doi: 10.1080/10810730.2023.2234866] [Medline: 37434532]
42. Goldsmith LP, Rowland-Pomp M, Hanson K, Deal A, Crawshaw AF, Hayward SE, et al. Use of social media platforms by migrant and ethnic minority populations during the COVID-19 pandemic: a systematic review. BMJ Open 2022;12(11):e061896 [FREE Full text] [doi: 10.1136/bmjopen-2022-061896] [Medline: 36396309]
43. Patil U, Kostareva U, Hadley M, Manganello JA, Okan O, Dadaczynski K, et al. Health literacy, digital health literacy, and COVID-19 pandemic attitudes and behaviors in U.S. college students: implications for interventions. Int J Environ Res Public Health 2021;18(6):3301 [FREE Full text] [doi: 10.3390/ijerph18063301] [Medline: 33806763]
44. Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. J Med Syst 2024;48(1):22 [FREE Full text] [doi: 10.1007/s10916-024-02045-3] [Medline: 38366043]

Abbreviations

AI: artificial intelligence
API: application programming interface
GQS: Global Quality Scale
LLM: large language model
PEMAT AV: Patient Education Materials Assessment Tool for Audiovisual Materials

Edited by E Lee; submitted 21.May.2025; peer-reviewed by KH Lin, C Baur, Y Zhang, J Antoun; comments to author 18.Aug.2025; revised version received 30.Dec.2025; accepted 09.Jan.2026; published 09.Feb.2026.

Please cite as:
Komsany A, Al Zoubi O, Sebaaly L, Harrison G, Soroka O, ElKefi S, Scales D, Phillips E, Pinheiro LC, Ismail I, Chebli P
Leveraging AI for Analysis of Digital Health Information on Cancer Prevention Among Arab Youth and Adults: Content Analysis
JMIR Infodemiology 2026;6:e77888
URL: https://infodemiology.jmir.org/2026/1/e77888
doi:10.2196/77888
PMID:

Original Paper

Health Data for Linguistic Minority Group Research in Canada: Proof-of-Concept Centralized Health Care Metadata Repository Development and Usability Study

Vincent Martin-Schreiber[1*], BScN, MEng; Cayden Peixoto[2*], HBSc, MSc; Ricardo Batista[3], MSc, MD, PhD; Christopher Belanger[4], HBSc, MBA, PhD; Peter Tanuseputro[5], HBSc, MSc, MD; Amy T Hsu[6], PhD; Lise M Bjerre[2,7,8], MDCM, PhD

[1]Faculty of Health Sciences, University of Ottawa, Ottawa, ON, Canada

[2]School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

[3]Akausivik Inuit Family Health Team, Ottawa, ON, Canada

[4]Telfer School of Management, University of Ottawa, Ottawa, ON, Canada

[5]Department of Family Medicine and Primary Care, University of Hong Kong, Hong Kong, China (Hong Kong)

[6]Bruyère Health Research Institute, Ottawa, ON, Canada

[7]Institut du Savoir Montfort, Ottawa, ON, Canada

[8]Department of Family Medicine, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

[*]these authors contributed equally

Corresponding Author:
Lise M Bjerre, MDCM, PhD
Department of Family Medicine
Faculty of Medicine
University of Ottawa
600 Peter Morand Crescent
Suite 201
Ottawa, ON, K1G 5Z3
Canada
Phone: 1 613 562 5800 ext 2982
Email: lbjerre@uottawa.ca

Abstract

Background: Language barriers between Canadian patients and health care providers are associated with poorer health outcomes, including decreased patient safety and quality of care, misdiagnosis and longer treatment initiation times, and increased mortality. However, research exploring language as a social determinant of health is limited, as Canadian health data are scattered across many jurisdictions, each with its own policies and procedures. This fragmentation makes it difficult for researchers to identify, locate, and use existing data. This paper presents the results of a pilot study that attempts to address this gap by creating a metadata repository (MDR) to act as a central source of information about what data are available at which data holdings across Canada.

Objective: This project aimed to (1) create a proof-of-concept MDR for Canadian health data at the variable level; (2) identify and label language-related variables existing within the MDR data; and (3) develop an interactive, public-facing web application to let users browse and search the MDR.

Methods: Metadata were collected from 5 Canadian health data sources, including 4 provincial data holdings and 1 national survey, and pooled to create a data repository. Then, we performed bottom-up labeling of language-related variables within the pooled metadata by first using a search string algorithm across all variable labels, names, and definitions and then consensus screening these variables using a derived, standardized definition of language or linguistic variables. Using the *Shiny* web framework in R, we then developed an openly accessible web application to allow users to search the proof-of-concept MDR.

Results: A total of 850,343 variables were collected and included in the repository, with most coming from Ontario (n=712,037, 83.7%) and Manitoba (n=97,051, 11.4%) provincial data holdings. Among all variables in the repository, 213,696 (25.1%) were confirmed to be language related.

Conclusions: Developing a national MDR would be a transformative opportunity for Canadian researchers to leverage the full scope of Canadian health administrative data. Although a top-down approach with consistent engagement of and collaboration between provincial data holdings and federal data agencies is ideal to develop a national MDR, this study demonstrates the feasibility of a bottom-up approach in contributing to this overarching goal.

(JMIR Infodemiology 2026;6:e77242) doi:10.2196/77242

KEYWORDS

metadata; metadata repository; variables; language; linguistic

Introduction

Background

Canada's publicly funded health care system generates a vast amount of data covering factors as wide ranging as pharmacy or prescription records, laboratory results, and health care services [1,2]. These data hold immense potential for health care research and for health policy and planning. However, because health care is administered differently across each of Canada's provinces and territories, the data are scattered across a large number of agencies and institutions, each with its own data policies and procedures [3]. This makes it difficult for researchers to access Canada's provincial health data and also creates a more fundamental problem—it is often difficult for researchers to even discover what types of data are available, where they are held, and how to access them. This fragmentation has contributed to significant differences in the availability and accessibility of administrative and other health data across provinces, posing a major challenge for interprovincial or pan-Canadian health care research. This "data fragmentation" can create particular problems for health care research related to patient and health care provider language abilities.

Language as a social determinant of health is an important and emerging topic in health research [4,5], and language barriers between Canadian patients and health care providers are associated with misdiagnosis and longer treatment initiation times [6]; negative experiences for patients [7,8] and physicians [9,10]; and, in hospital settings, decreased patient safety and quality of care [11,12] as well as increased mortality [13]. This issue is of specific concern in Canada, an officially bilingual country in which 76.1% of the population are native English speakers, 22% are native French speakers, and 18% are bilingual [14]. Although French speakers and English speakers can be found across the country, most French speakers live in the provinces of Quebec and New Brunswick. Despite the importance of language-related health research, the data fragmentation described previously makes it difficult and time consuming to even discover what language-related data are available, let alone access and analyze them. This paper presents the results of a pilot study that attempts to bridge this gap by creating a "metadata repository" (MDR) to serve as a central source of information about which data are available at which locations across Canada.

Metadata can be defined as "data about data" [15], and for this project, we sought to create a repository of variable-level metadata. In this context, variable-level metadata include information such as the institution holding the variable, the larger collection or "library" to which it belongs, and a plaintext description. To help illustrate the utility of MDR in light of Canada's bilingual health care context, we put a special focus on identifying language-related variables. In addition to our final metadata dataset, we also created an interactive public-facing web application to let users browse and search the repository.

We discuss the current state of health data and metadata management in Canada and outline the principles and scope guiding our pilot project subsequently.

Current Initiatives in Canada

There are currently 2 main health metadata initiatives in Canada: the Health Data Research Network (HDRN) Canada's Data Access Support Hub (DASH) and the Strategy for Patient-Oriented Research (SPOR) Canadian Data Platform (CDP). The HDRN is a pan-Canadian network of health data–holding organizations, and it established DASH [16] to guide researchers and streamline access to data held by its members. However, DASH only helps researchers access data housed at member organizations of HDRN Canada, and its services are not free to use.

The SPOR CDP, announced in 2019 by Canada's Ministry of Health, is intended to function as a single portal for researchers to request access to administrative, clinical, and social data from sources across the country [17]. To achieve this goal, the SPOR CDP aims to harmonize and validate definitions for key analytic variables (eg, chronic diseases) while expanding the sources, types, and linkages of data available to researchers (eg, social data). Standardizing data definitions allows information exchanged between data holdings to be equally understood by all parties, a concept known as semantic interoperability [18]. Semantic interoperability is especially important as it allows researchers to combine datasets. Canada is known to lag in health data interoperability [19-21], and the development of metadata standards, a set of guidelines that establish a common way of structuring and understanding data [15], would be very helpful. However, the CDP platform was originally announced as a 7-year initiative and is still ongoing as of 2026.

In addition to larger metadata projects, some data-holding organizations also have public-facing websites that allow users to search their metadata. For example, the Institute for Clinical Evaluative Sciences (ICES) in Ontario provides a publicly accessible data dictionary of their metadata that is searchable at the variable level [22]. Although such resources can be helpful, they lead to the problem of data fragmentation described previously, as researchers must visit each organization's website and consolidate results themselves.

Although there are clear use cases for larger projects such as DASH or the CDP and smaller, institution-level metadata websites, they do not offer a free-to-use and up-to-date repository of health metadata from across Canada. The goal of this study is to take the initial steps toward bridging this gap.

A National MDR: Pilot Project Principles

In this study, we were guided by 2 sets of principles: the principles of findability, accessibility, interoperability, and reusability (FAIR) data stewardship [23] and a bottom-up principle of researcher-driven development.

The FAIR principles were developed by Wilkinson et al [23] to address the challenges in managing large amounts of data. The FAIR principles stipulate that both data and metadata should be findable, accessible, interoperable, and reusable by researchers [23,24]. Clearly, the fragmented landscape of Canadian administrative health data does not adhere to FAIR principles in this sense, which creates what we view as unnecessary delays and roadblocks to potentially life-saving research.

We also postulate that there is a useful role for researchers to play in creating a pragmatic and useful national MDR within the current Canadian health data landscape. Given the size and complexity of administrative health databases and the dappled policy environment governing data access across Canada, creating an MDR through the top-down approach at the organizational level would take a large degree of coordination, political will, and resources to harmonize data selection, definition, collection, and sharing procedures across all provincial and territorial health data holdings. Although an MDR built through top-down standardization would be ideal, there is no guarantee that one will be available in Canada soon.

According to the Public Health Agency of Canada, federal, provincial, and territorial governments are currently working to improve the sharing of public health information [25]. However, a data-sharing agreement between these governments is not expected until the end of 2026, with bilateral agreements to follow and then a lengthy process of harmonizing definitions and processes across the data holdings. In the meantime, a simpler solution built by and for researchers has the potential to provide value now.

Methods

Data Sources and Data Collection

To ensure that our proof-of-concept MDR is robust and inclusive, we aimed to include metadata from a variety of national and provincial administrative health data sources. Administrators and data custodians at national and provincial data holdings (Table 1) were contacted via email between January 2023 and September 2023 to request access to the metadata from all held administrative health datasets, ideally in a raw data format such as CSV. Among the data custodians contacted, metadata were provided by or accessible from the ICES [26], the Manitoba Centre for Health Policy (MCHP) [27], the Institut de la statistique du Québec [28], and the New Brunswick Institute for Research, Data and Training [29]. We also obtained and included metadata from the Canadian Longitudinal Study on Aging [30].

Metadata files were obtained from MCHP and the New Brunswick Institute for Research, Data and Training. For the other 4 data holdings, data scraping [31] was performed by a member of the research team (VM-S) to extract metadata from publicly available online sources and data dictionaries. Detailed explanations of how data were collected from each data holding are provided in Multimedia Appendix 1. Once the metadata from the 5 included data holdings were pooled into a single CSV file, the metadata were organized according to commonly reported data elements across sources, including data holding, dataset name, dates available, variable label, variable name, and variable definition, as reported by the respective data source.

Table 1. Data dictionary availability from administrative health data custodians by province.

Province	Data custodian	Publicly accessible data dictionary or catalog
Alberta	Alberta Health	Yes[a]
BC[b]	Population Data BC	Yes
Manitoba	Manitoba Centre for Health Policy	Yes
New Brunswick	New Brunswick Institute for Research, Data and Training	Yes
Newfoundland and Labrador	Newfoundland and Labrador Centre for Health Information	Yes
Nova Scotia	Health Data Nova Scotia	Yes[a]
Ontario	Institute for Clinical Evaluative Sciences	Yes
PEI[c]	Health PEI	No
Quebec	Régie de l’assurance maladie du Québec	No
Saskatchewan	eHealth Saskatchewan	No

[a]Available only upon request.

[b]BC: British Columbia.

[c]PEI: Prince Edward Island.

Data Labeling

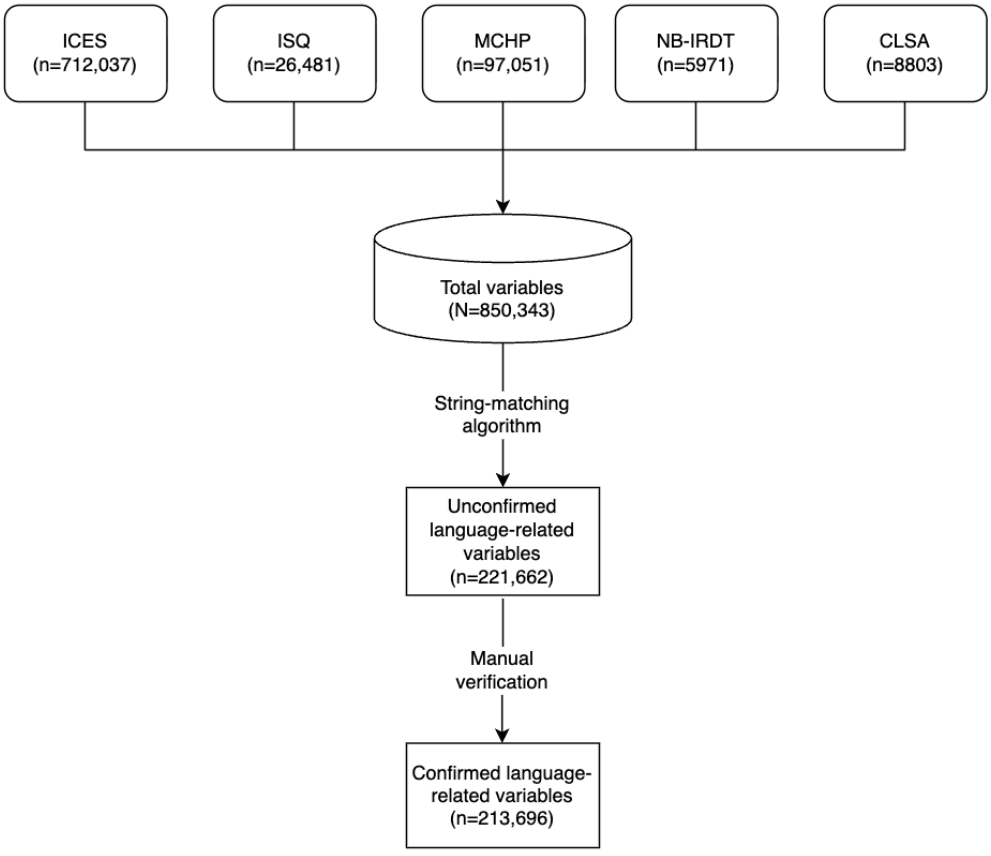
Data labeling (or tagging) is the common process of assigning one or more descriptive tags or labels to a dataset [32], which can make it easier to search and filter results while enabling other uses of the data (eg, machine learning) [33]. To provide an example of searchability in our proof-of-concept MDR, we identified potential language-related variables. We used a naive string-searching algorithm, which works by checking for the occurrence of a pattern (or string) at every possible position in the text [34]. Given Canada's status as a bilingual English-speaking and French-speaking country, we identified potentially linguistic variables as those matching any of the following text strings: "french," "english," "lang," "spoken," "speak," "ling," and "franc."

From here, 2 members of the research team (VM-S and CP) independently reviewed 9.9% (84,068/850,343) of the overall variable names and definitions in the dataset (these were taken from the list of potential language-related variables; Figure 1) to agree on the criteria to define what a language variable is, placing higher value on the most common definitions. The 2 researchers then met to reach a consensus on the standardized definition for tagging language-related variables within the proof-of-concept MDR: "any variable that directly or indirectly provides information regarding the linguistic characteristics of an individual, a health professional, or an organization." This definition aimed to be extremely broad to be able to accommodate any form of research method, including Bayesian statistical approaches.

Both researchers then independently screened all previously identified variables, including variable names and definitions, to identify all language-related variables according to our standardized definition. Screening results were then compared to ensure consensus in the labeled variables between the 2 researchers. Any conflicts in the identification of language-related variables or in the application of the standardized tagging definition were resolved via conversation with a third member of the research team (LMB).

To quantitatively assess the reliability of this screening process, we calculated interrater reliability using Cohen κ [35], which measures the consistency with which both researchers independently applied the standardized definition, accounting for agreement that would be expected by chance. This approach is a standard practice in systematic reviews and content analysis methodologies where operational definitions are developed through iterative refinement and consensus-building discussion [36]. To reduce potential bias, we calculated Cohen κ on 137,594 of the 221,662 (62.0%) variables screened after the standardized definition was established (Figure 1), excluding the 84,068 (37.9%) variables from the initial screening phase used for definition development.

Figure 1. Flowchart for identification of language-related variables from health care data holdings. CLSA: Canadian Longitudinal Study on Aging; ICES: Institute for Clinical Evaluative Sciences; ISQ: Institut de la statistique du Québec; MCHP: Manitoba Centre for Health Policy; NB-IRDT: New Brunswick Institute for Research, Data and Training.



Ethical Considerations

All data used in this study were limited to publicly available metadata, which posed no privacy risk or potential for harm. No personal health information, patient data, or confidential research data were accessed. The collected metadata consisted solely of variable names, descriptions, dataset structures, and data availability information—content that data custodians have chosen to make publicly available to facilitate research discovery and data access applications. For these reasons, approval from a research ethics board was neither required nor sought.

Results

Overview

Across the 5 included data sources, metadata from a total of 850,343 variables were collected and included in our repository. The number of metadata variables collected from each data holding is presented in Table 2. Among the data holdings, the ICES (712,037/850,343, 83.7%) and MCHP (n=97,051, 11.4%) data holdings contained the most variables.

Among the initial 850,343 variables in our repository, 221,662 (26.1%) potential or unconfirmed language-related variables were identified by using a search string algorithm across variable labels, names, and definitions. Consensus screening of these variables using a derived, standardized definition of language or linguistic variables identified 213,696 (25.1%) confirmed language-related variables in our repository (Figure 1).

Interrater reliability for the independent screening process was assessed using observed percent agreement and Cohen κ. The 2 researchers initially agreed on 96.3% (132,538/137,594) of these postdefinition variables, with 5056 (3.7%) disagreements that were subsequently resolved through consensus-building discussion. The calculated Cohen κ was 0.621 (95% CI 0.611-0.632), indicating substantial agreement between the 2 researchers.

Table 2. Number of metadata variables included in the proof-of-concept metadata repository by data holding (N=850,343).

Data holding	Variables, n (%)
Institute for Clinical Evaluative Sciences	712,037 (83.7)
Institut de la statistique du Québec	26,481 (3.1)
Manitoba Centre for Health Policy	97,051 (11.4)
New Brunswick Institute for Research, Data and Training	5971 (0.7)
Canadian Longitudinal Study on Aging	8803 (1.0)

Creating a Usable Proof-of-Concept Web Interface

To facilitate exploration and use of the MDR, we developed a prototype web application that allows users to browse and search the MDR over the internet [37]. Although data management frameworks exist, such as the DataHub Project [38] and the Comprehensive Knowledge Archive Network (CKAN) [39], we developed our application in R (version 4.3.1; R Foundation for Statistical Computing) using *Shiny* [40]. *Shiny* is an open-source R package that makes it easy to build interactive web applications directly using R, a programming language widely used for statistical computing and graphics. *Shiny* was chosen due to its relative simplicity compared to other web development frameworks and the research team's familiarity with the R programming language. The user interface of the *Shiny* app was designed with user-friendliness and functionality in mind, and it allows users to search by keyword, filter by data properties (eg, data holding and linguistic properties), and browse through paginated results. The *Shiny* application was built into a Docker image and hosted on a public platform-as-a-service provider.

Discussion

Principal Findings

Canada's health data are scattered across many organizations and jurisdictions, each with its own policies and procedures, making it difficult for researchers to identify, locate, and use existing data [23,41]. To address this gap, we developed a proof-of-concept MDR containing metadata for more than 850,000 variables from 5 different Canadian data holdings and performed bottom-up labeling of 213,696 (25.1%) of the 850,343 language-related variables within the repository to help researchers easily identify language-related data within the vast landscape of Canadian health data. We also developed an openly accessible web application to allow users to search for the MDR [37].

Building a Bottom-Up MDR: Lessons Learned

Our pilot project demonstrated the feasibility of a bottom-up approach to building an MDR for Canadian health data, but we learned several important lessons that we summarize here. First, complex, manual effort was required to collect (or "scrape") data that are publicly available on the internet. Web scraping is very fast when it works, but each data source needs a bespoke approach. Some websites are straightforward to scrape (eg, those that use backend application programming interfaces) that can be queried directly), but others use an architecture that is not well suited to automatic data collection (eg, those that require repeated form submissions or client-side JavaScript). In addition, the scraping logic is custom-built to each website's design at that moment in time, and if repositories update their websites, the scraping code will need to be updated as well.

Second, there is a need for robust internal data management practices when developing an MDR. We initially prioritized simplicity as well as data portability and transparency; therefore, we stored our data in plaintext CSV files. However, as we collected more data, we were surprised by the size of our final dataset, at a little more than 1 GB. Although this is small

compared to many geospatial or genetic datasets, files of this size are unwieldy to work with, since common office software, such as Microsoft Excel, may not be able to load all variables and can be slow and difficult to transfer to others. For any similar projects, we suggest that a simple data-storage format, such as CSV files, is appropriate for initial feasibility studies, but the project should move quickly to a more sophisticated centralized data-storage solution (eg, a database or a large-file storage solution with version tracking) once feasibility has been established.

Finally, we learned that *Shiny* has several limitations that make it ill-suited for public-facing web applications with datasets this large. *Shiny* creates a new R session for each user and loads the entire dataset into server memory. For a typical dataset measured in KBs or MBs, the overhead is negligible; however, since our data are approximately 1 GB, our application runs out of memory and crashes with more than a few concurrent users. So, although *Shiny* was indispensable to us for rapid prototyping on a local computer, for production deployments, we suggest a different framework in which the data are stored in a single database and queried as needed, as opposed to the server loading a new in-memory copy of the dataset for each user. The user interface could be written using any web development framework (eg, Phoenix and React) and the open-source database software such as PostgreSQL, which is commonly used in large commercial and government projects, would be capable of handling queries on a million-row dataset with millisecond-level response times [42]. Direct access to the application programming interface could also be added, but implementation details of this potential future project are outside the scope of this paper.

Limitations

Although our proof of concept provides a working example of a bottom-up labeled MDR, our methodology is not without limitations. For our initial screening of language variables, we used a search string algorithm to first identify potential language-related variables within all datasets in the proof-of-concept MDR. This search string may not have been exhaustive and could have missed potential language-related variables within the included datasets. Moreover, to best assess the accuracy of our algorithm, it would need to be tested against manually screened datasets as a gold standard for our definition. Although this process was considered too time consuming for the scope of this proof-of-concept project, given the size of the datasets used, it would allow us to evaluate measures such as sensitivity and specificity.

Finally, because variables in our repository were web scraped from various data sources, our repository reflects what variables were available at the point in time of data scraping and would require a repeat of the scraping, screening, and labeling process to update the repository as it is. In addition, there may have been additional metadata in the data holdings that were not made publicly available and therefore were not scrapable, meaning our proof-of-concept MDR may not be exhaustive of variables from the included data holdings. Nonetheless, without top-down policies and procedures in place to allow for easy data collection and labeling processes across Canada, our language-variable data labeling provides a working example of how bottom-up data labeling can be performed by researchers.

Future Directions

Although currently in a beta version, we have plans to expand the MDR to include variables from additional Canadian administrative health data holdings, such as Population Data British Columbia [43], and data from Statistics Canada [44]. Moreover, additional variable tagging can be performed to identify sociodemographic variable types within all included datasets for research purposes, such as sex, gender, race, ethnicity, income, and immigration status. Regarding language-related variables specifically, subtagging can be performed for more specific variable definitions [45], including knowledge of official languages (French or English), variables indicating first language or mother tongue, or variables related to patient–health care provider language concordance.

We also intend to develop a new MDR web application that overcomes the limitations of our *Shiny* app by using a backend database, so that the entire dataset does not need to be loaded into server memory repeatedly for each user.

Finally, we believe that creating a *top-down* national MDR is a worthy goal that should be pursued in tandem with *bottom-up* efforts such as ours. However, such a project would face a number of governance, legal, ethical, and administrative barriers and require a high degree of alignment across diverse organizations so as not to create numerous delays in the collection and integration of data from provincial and organizational data custodians [46,47]. In other words, an ideal top-down MDR will need intense collaboration between many organizations, and although this is beyond our power as individual researchers, we hope Canada's data custodians will rise to the challenge.

Conclusions

This paper addresses the need for a national MDR of administrative and other health data in Canada, underscoring how an MDR can address issues caused by data fragmentation and increase the FAIRness of health care data across the country. However, complex challenges hinder the development of a top-down health data MDR in Canada. We developed a proof-of-concept MDR of administrative health data from 5 different data sources and performed bottom-up labeling of language-related variables within the repository to help researchers easily identify language data in the vast landscape of Canadian health data. This MDR is publicly available online as a searchable data dictionary [37].

Our proof-of-concept MDR illustrates the methodological limitations of a bottom-up approach, which can be complementary and synergistic with but cannot replace top-down approaches to the development of such a repository. Engagement of and collaboration between provincial data holdings and federal data agencies are critical to ensuring a pan-Canadian MDR is comprehensive and can be kept up to date. A national MDR would make it simple and straightforward for Canadian researchers to leverage the full scope of Canadian health data, and open opportunities for new studies as researchers discover datasets previously unknown to them. We

believe that this could be transformative, and we hope this pilot project demonstrates the feasibility of a bottom-up approach in contributing toward this overarching goal.

Acknowledgments

The authors acknowledge the contributions of the following steering committee members and other contributors who, while not fulfilling all authorship criteria, have generously provided their expertise and contributed to the success of this project: Marie-Hélène Chomienne, associate professor and research chair, Chaire de recherche en francophonie internationale et santé de l'immigrant ou du réfugié d'Afrique francophone subsaharienne, University of Ottawa; Jan Warnke, data model analyst, l'Hôpital Jeffery Hale; Claire Kendall, senior scientist, Bruyère Health Research Institute; Cynthia Kendell, assistant professor and research implementation scientist, Department of Medicine, Dalhousie University; and Louise Bouchard, co-director, Réseau Francophonie Observatoire de Recherche Collaborative en Études sur la Santé et les services en contexte minoritaire.

Funding

This work was funded through a Canadian Institutes of Health Research Catalyst Grant: Official Language Minority Communities in Health Research (grant 472426; principal investigator LMB).

Data Availability

The data for this study are available for browsing [37], and the full dataset is available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data collection methods and compliance by source.
[DOCX File , 22 KB - infodemiology_v6i1e77242_app1.docx]

References

1. Cadarette SM, Wong L. An introduction to health care administrative data. Can J Hosp Pharm 2015 Jun 25;68(3):232-237 [FREE Full text] [doi: 10.4212/cjhp.v68i3.1457] [Medline: 26157185]
2. Lucyk K, Lu M, Sajobi T, Quan H. Administrative health data in Canada: lessons from history. BMC Med Inform Decis Mak 2015 Aug 19;15(1):69 [FREE Full text] [doi: 10.1186/s12911-015-0196-9] [Medline: 26286712]
3. Kendell C, Levy AR, Porter G, Gibson E, Urquhart R. Factors affecting access to administrative health data for research in Canada: a study protocol. Int J Popul Data Sci 2021 Sep 23;6(1):1653 [FREE Full text] [doi: 10.23889/ijpds.v6i1.1653] [Medline: 34632104]
4. Mansoor Y, Wong T, Comeau JL. Language: the ignored determinant of health. Paediatr Child Health 2024 Sep 12;29(3):168-170. [doi: 10.1093/pch/pxad066] [Medline: 38827371]
5. Batista R, Reaume M, Roberts R, Seale E, Rhodes E, Sucha E, et al. Prevalence and patterns of multimorbidity among linguistic groups of patients receiving home care in Ontario: a retrospective cohort study. BMC Geriatr 2023 Nov 09;23(1):725 [FREE Full text] [doi: 10.1186/s12877-023-04267-5] [Medline: 37946126]
6. de Moissac D, Bowen S. Impact of language barriers on quality of care and patient safety for official language minority francophones in Canada. J Patient Exp 2019 Mar 18;6(1):24-32 [FREE Full text] [doi: 10.1177/2374373518769008] [Medline: 31236448]
7. Jutras C, Gauthier AP, Timony PE, Côté D, Kpazaï G. Expérience de francophones en Ontario chez leur médecin de famille: concordance et discordance linguistique. Divers Res Health J 2020 Mar;3:12-33 [FREE Full text] [doi: 10.28984/drhj.v3i0.310]
8. Bowen S. The impact of language barriers on patient safety and quality of care. Société Santé en français. 2015 Aug. URL: https://www.santefrancais.ca/wp-content/uploads/2018/11/SSF-Bowen-S.-Language-Barriers-Study-1.pdf [accessed 2025-12-15]
9. Timony PE, Gauthier AP, Serresse S, Goodale N, Prpic J. Barriers to offering French language physician services in rural and northern Ontario. Rural Remote Health 2016;16(2):3805 [FREE Full text] [Medline: 27316568]
10. Gauthier AP, Timony PE, Serresse S, Goodale N, Prpic J. Strategies for improved French-language health services: perspectives of family physicians in northeastern Ontario. Can Fam Physician 2015 Aug;61(8):e382-e390 [FREE Full text] [Medline: 26505060]

11. Reaume M, Batista R, Talarico R, Guerin E, Rhodes E, Carson S, et al. In-hospital patient harm across linguistic groups: a retrospective cohort study of home care recipients. J Patient Saf 2022 Jan 01;18(1):e196-e204. [doi: 10.1097/PTS.0000000000000726] [Medline: 32433437]
12. Reaume M, Batista R, Talarico R, Rhodes E, Guerin E, Carson S, et al. The impact of hospital language on the rate of in-hospital harm. A retrospective cohort study of home care recipients in Ontario, Canada. BMC Health Serv Res 2020 Apr 21;20(1):340 [FREE Full text] [doi: 10.1186/s12913-020-05213-6] [Medline: 32316965]
13. Seale E, Reaume M, Batista R, Eddeen AB, Roberts R, Rhodes E, et al. Patient-physician language concordance and quality and safety outcomes among frail home care recipients admitted to hospital in Ontario, Canada. CMAJ 2022 Jul 11;194(26):E899-E908 [FREE Full text] [doi: 10.1503/cmaj.212155] [Medline: 35817434]
14. Statistics on official languages in Canada. Government of Canada. URL: https://www.canada.ca/en/canadian-heritage/services/official-languages-bilingualism/publications/statistics.html [accessed 2025-12-15]
15. Ulrich H, Kock-Schoppenhauer AK, Deppenwiese N, Gött R, Kern J, Lablans M, et al. Understanding the nature of metadata: systematic review. J Med Internet Res 2022 Jan 11;24(1):e25440 [FREE Full text] [doi: 10.2196/25440] [Medline: 35014967]
16. About DASH. Health Data Research Network Canada. URL: https://www.hdrn.ca/dash/about-dash/ [accessed 2025-12-15]
17. Health minister announces launch of SPOR Canadian Data Platform. Health Data Research Network Canada. URL: https://www.hdrn.ca/en/news/formal-launch-of-the-spor-canadian-data-platform/ [accessed 2025-12-15]
18. de Mello BH, Rigo SJ, da Costa CA, da Rosa Righi R, Donida B, Bez MR, et al. Semantic interoperability in health records standards: a systematic literature review. Health Technol (Berl) 2022;12(2):255-272 [FREE Full text] [doi: 10.1007/s12553-022-00639-w] [Medline: 35103230]
19. Read KB, Gibson G, Leahey A, Peterson L, Rutle S, Shi J, et al. Identifying metadata commonalities across restricted health data sources: a mixed methods study exploring how to improve the discovery of and access to restricted datasets. J Escience Librariansh 2024;13(2):e907 [FREE Full text] [doi: 10.7191/jeslib.907]
20. Expert Advisory Group. Pan-Canadian health data strategy: toward a world-class health data system. Public Health Agency of Canada. 2022. URL: https://tinyurl.com/ycfb2r8z [accessed 2026-01-23]
21. Affleck E, Sutherland E, Lindeman C, Golonka R, Price T, Murphy T, et al. Human factor health data interoperability. Healthc Pap 2024 Jan 31;21(4):47-55. [doi: 10.12927/hcpap.2024.27272] [Medline: 38482657]
22. Data dictionary. Institute for Clinical Evaluative Sciences (ICES). URL: https://datadictionary.ices.on.ca/Applications/datadictionary/Default.aspx [accessed 2025-12-15]
23. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3(1):160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
24. Boeckhout M, Zielhuis GA, Bredenoord AL. The FAIR guiding principles for data stewardship: fair enough? Eur J Hum Genet 2018 Jul 17;26(7):931-936 [FREE Full text] [doi: 10.1038/s41431-018-0160-0] [Medline: 29777206]
25. Modernizing public health information sharing. Government of Canada. URL: https://www.canada.ca/en/public-health/services/data/modernizing-information-sharing.html [accessed 2025-12-15]
26. Institute for Clinical Evaluative Sciences homepage. Institute for Clinical Evaluative Sciences (ICES). URL: https://www.ices.on.ca/ [accessed 2025-12-15]
27. Manitoba Centre for Health Policy. University of Manitoba. URL: https://umanitoba.ca/manitoba-centre-for-health-policy/ [accessed 2025-12-15]
28. Institut de la statistique du Québec homepage. Institut de la statistique du Québec. URL: https://statistique.quebec.ca/en [accessed 2025-12-15]
29. DataNB. University of New Brunswick. URL: https://www.unb.ca/nbirdt/ [accessed 2025-12-15]
30. CLSA homepage. Canadian Longitudinal Study of Aging. URL: https://www.clsa-elcv.ca/ [accessed 2025-12-15]
31. Web scraping. Statistics Canada. URL: https://www.statcan.gc.ca/en/our-data/where/web-scraping [accessed 2025-12-15]
32. Mahalle PN, Shinde GR, Ingle YS, Wasatkar NN. Data Centric Artificial Intelligence: A Beginner's Guide. Cham, Switzerland: Springer; 2023.
33. Paun S, Artstein R, Poesio M. Statistical Methods for Annotation Analysis. Cham, Switzerland: Springer; 2022.
34. Abdeen RA. An algorithm for string searching. Int J Comput Appl 2019 Oct 17;177(10):17-22 [FREE Full text] [doi: 10.5120/ijca2019919484]
35. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
36. Belur J, Tompson L, Thornton A, Simon M. Interrater reliability in systematic review methodology: exploring variation in coder decision-making. Sociol Methods Res 2018 Sep 24;50(2):837-865. [doi: 10.1177/0049124118799372]
37. CLOSM - data catalogue (beta version). Health Data Dictionary. 2024. URL: https://healthdatadictionary.ca/ [accessed 2025-12-15]
38. Datahub-project. GitHub. URL: https://github.com/datahub-project/datahub [accessed 2025-12-15]
39. CKAN. GitHub. URL: https://github.com/ckan [accessed 2025-12-15]
40. Rstudio/shiny. GitHub. URL: https://github.com/rstudio/shiny [accessed 2025-12-15]

41. Neumann J. FAIR data infrastructure. Adv Biochem Eng Biotechnol 2022;182:195-207. [doi: 10.1007/10_2021_193] [Medline: 35091812]
42. Salunke SV, Ouda A. A performance benchmark for the PostgreSQL and MySQL databases. Future Internet 2024 Oct 19;16(10):382. [doi: 10.3390/fi16100382]
43. Population Data BC homepage. Population Data BC. URL: https://www.popdata.bc.ca/ [accessed 2025-12-15]
44. Data. Statistics Canada. URL: https://www150.statcan.gc.ca/n1/en/type/data [accessed 2025-12-15]
45. Batista R, Hsu AT, Bouchard L, Reaume M, Rhodes E, Sucha E, et al. Ascertaining the Francophone population in Ontario: validating the language variable in health data. BMC Med Res Methodol 2024 Apr 27;24(1):98 [FREE Full text] [doi: 10.1186/s12874-024-02220-7] [Medline: 38678174]
46. Katz A, Enns J, Wong ST, Williamson T, Singer A, McGrail K, et al. Challenges associated with cross-jurisdictional analyses using administrative health data and primary care electronic medical records in Canada. Int J Popul Data Sci 2018 Oct 05;3(3):437 [FREE Full text] [doi: 10.23889/ijpds.v3i3.437] [Medline: 34095523]
47. Become an ICES Scientist. Institute for Clinical Evaluative Sciences (ICES). URL: https://www.ices.on.ca/join-our-research-community/become-an-ices-scientist/ [accessed 2025-12-15]

Abbreviations

CDP: Canadian Data Platform
CKAN: Comprehensive Knowledge Archive Network
DASH: Data Access Support Hub
FAIR: findability, accessibility, interoperability, and reusability
HDRN: Health Data Research Network
ICES: Institute for Clinical Evaluative Sciences
MCHP: Manitoba Centre for Health Policy
MDR: metadata repository
SPOR: Strategy for Patient-Oriented Research

Edited by I Brooks; submitted 09.May.2025; peer-reviewed by H Spechbach, B Cahill, M Al Zoubi; comments to author 29.Oct.2025; revised version received 19.Dec.2025; accepted 14.Jan.2026; published 09.Feb.2026.

Please cite as:
Martin-Schreiber V, Peixoto C, Batista R, Belanger C, Tanuseputro P, Hsu AT, Bjerre LM
Health Data for Linguistic Minority Group Research in Canada: Proof-of-Concept Centralized Health Care Metadata Repository Development and Usability Study
JMIR Infodemiology 2026;6:e77242
URL: https://infodemiology.jmir.org/2026/1/e77242
doi:10.2196/77242
PMID:41543876

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org