Original Paper

Predicting Openness of Communication in Families With Hereditary Breast and Ovarian Cancer Syndrome: Natural Language Processing Analysis

Vasiliki Baroutsou[1], MSc; Rodrigo Cerqueira Gonzalez Pena[2], PhD; Reka Schweighoffer[1], PhD; Maria Caiata-Zufferey[3], PhD; Sue Kim[4], PhD; Sharlene Hesse-Biber[5], PhD; Florina M Ciorba[6], PhD; Gerhard Lauer[7], PhD; Maria Katapodi[1], PhD; CASCADE Consortium[8]

[1]Department of Clinical Research, University of Basel, Basel, Switzerland

[2]Center for Data Analytics, University of Basel, Basel, Switzerland

[3]Competence Centre for Healthcare Practices and Policies, Department of Business Economics, Health and Social Care, University of Applied Sciences and Arts of Southern Switzerland, Manno, Switzerland

[4]College of Nursing, Yonsei University, Seoul, Republic of Korea

[5]Department of Sociology, Boston College, Chestnut Hill, MA, United States

[6]Department of Mathematics and Computer Science, University of Basel, Basel, Switzerland

[7]Gutenberg-Institut für Weltliteratur und schriftorientierte Medien, Abteilung Buchwissenschaft Johannes Gutenberg, Universität Mainz Philosophicum, Mainz, Germany

[8]See Acknowledgments

Corresponding Author:
Maria Katapodi, PhD
Department of Clinical Research
University of Basel
Missionstrasse 64
Basel, 4055
Switzerland
Phone: 41 612070430
Email: maria.katapodi@unibas.ch

Abstract

Background: In health care research, patient-reported opinions are a critical element of personalized medicine and contribute to optimal health care delivery. The importance of integrating natural language processing (NLP) methods to extract patient-reported opinions has been gradually acknowledged over the past years. One form of NLP is sentiment analysis, which extracts and analyses information by detecting feelings (thoughts, emotions, attitudes, etc) behind words. Sentiment analysis has become particularly popular following the rise of digital interactions. However, NLP and sentiment analysis in the context of intrafamilial communication for genetic cancer risk is still unexplored. Due to privacy laws, intrafamilial communication is the main avenue to inform at-risk relatives about the pathogenic variant and the possibility of increased cancer risk.

Objective: The study examined the role of sentiment in predicting openness of intrafamilial communication about genetic cancer risk associated with hereditary breast and ovarian cancer (HBOC) syndrome.

Methods: We used narratives derived from 53 in-depth interviews with individuals from families that harbor pathogenic variants associated with HBOC: first, to quantify openness of communication about cancer risk, and second, to examine the role of sentiment in predicting openness of communication. The interviews were conducted between 2019 and 2021 in Switzerland and South Korea using the same interview guide. We used NLP to extract and quantify textual features to construct a handcrafted lexicon about interpersonal communication of genetic testing results and cancer risk associated with HBOC. Moreover, we examined the role of sentiment in predicting openness of communication using a stepwise linear regression model. To test model accuracy, we used a split-validation set. We measured the performance of the training and testing model using area under the curve, sensitivity, specificity, and root mean square error.

Results: Higher "openness of communication" scores were associated with higher overall net sentiment score of the narrative, higher fear, being single, having nonacademic education, and higher informational support within the family. Our results

demonstrate that NLP was highly effective in analyzing unstructured texts from individuals of different cultural and linguistic backgrounds and could also reliably predict a measure of "openness of communication" (area under the curve=0.72) in the context of genetic cancer risk associated with HBOC.

Conclusions: Our study showed that NLP can facilitate assessment of openness of communication in individuals carrying a pathogenic variant associated with HBOC. Findings provided promising evidence that various features from narratives such as sentiment and fear are important predictors of interpersonal communication and self-disclosure in this context. Our approach is promising and can be expanded in the field of personalized medicine and technology-mediated communication.

(JMIR Form Res 2023;7:e38399) doi: 10.2196/38399

KEYWORDS

cascade testing; dictionary-based approach; family communication; hereditary breast and ovarian cancer; HBOC; sentiment analysis; text mining; natural language processing; cancer; hereditary

Introduction

Natural language processing (NLP) is a computer-assisted analytical approach for automatically evaluating and interpreting human language by extracting meaningful insights from textual data sets [1-3]. NLP has been broadly used in various fields in the recent past, for example, in financial and business marketing, education, and health care [4-8]. The typical applications of NLP include information extraction, sentiment and semantic analysis, text classification, and text summarization. Among the different NLP applications, sentiment analysis has become particularly popular in recent years following the rise of digital communication and social media [2,9]. Sentiment analysis aims to assess whether people's opinions, emotions, and attitudes toward a certain event or experience are positive, negative, or neutral [3,10,11] and generates valuable insights that lead to the improvement of a new service or product.

In health care–related studies, patient-reported insights are an essential component of personalized medicine and contribute to optimal health care delivery. Researchers have applied NLP to extract and analyze patient-reported insights from social media and for different topics, for example, social exchange patterns in web-based health platforms [12], needs of patients and caregivers in different disease entities [13], online support groups for patients with breast cancer [14], or awareness for Lynch syndrome (LS) [15]. A major limitation of this approach is that population characteristics (age, socioeconomic status, etc) are often unavailable, which limits the clinical applicability of findings and may create disparities either due to increased representation or lack thereof of certain population subgroups. Others have applied NLP to clinical notes originating from electronic medical records to describe patients' experiences with symptoms [16] or free-text data from patient surveys evaluating the quality of hospital services [17]. One limitation of this approach is the lack of depth in these data sources, either because they lack the patient's perspective or because the texts are limited in scope and volume. We identified only a few studies that applied NLP to unstructured narratives collected from in-depth interviews aiming to describe experiences with cancer ambulatory services [18] or to predict changes in substance use [19] and perceived loneliness among older adults [20].

NLP and sentiment analysis in the context of intrafamilial communication for genetic cancer risk is unexplored. Due to privacy laws, individuals carrying pathogenic variants in cancer-causing genes have a key role in disseminating information to relatives and in advocating for genetic testing [21]. This self-disclosure process is currently the main avenue to alert relatives to their own risk of carrying the pathogenic variant. Self-disclosure is a process of interpersonal communication by which one person reveals information about themselves to another person, or a small intimate group, for example, their family. The information exchange can be based on verbal and nonverbal cues and can be face to face or technology mediated. Most importantly, in addition to information exchange, self-disclosure can include thoughts, emotional experiences and feelings, aspirations, goals, fears, likes, and dislikes [22]. During self-disclosure, humans adjust and adapt their verbal and nonverbal communication, and messages are produced, interpreted, understood, or misunderstood [23,24]. Intrafamilial communication for genetic cancer risk may involve significant levels of uncertainty and potential conflicts since the meaning of self-disclosure about the cancer-causing variant can be shaped by opposing arguments and negative responses from others. Indeed, information exchange about genetic cancer risk may be easier with some family members or may present a particularly difficult moment with others [25,26].

Predicting openness of communication and examining the role of sentiment in intrafamilial communication of genetic cancer risk may be used to enrich message tailoring in technology-assisted interventions. In this study, we examined the role of sentiment in predicting openness of communication about genetic cancer risk associated with hereditary breast and ovarian cancer (HBOC) syndrome. HBOC is a hereditary cancer syndrome that affects both men and women and accounts for a significant number of different cancers, such as breast, ovarian, pancreatic, and prostate [27]. Sharing information about HBOC-causing pathogenic variants is a complex process of intrafamilial communication and a key element of public health interventions aiming to promote cascade testing of relatives and cancer prevention and control [28,29]. In this study, we used narrative data collected with in-depth interviews: first, to quantify openness of communication about HBOC cancer risk, and second, to examine the role of sentiment in predicting openness of communication.

Methods

Design, Population, Settings, and Procedures

This analysis is part of a larger ongoing study, the Swiss CASCADE cohort, which follows adult (aged ≥18 years) men and women from families that harbor pathogenic variants associated with HBOC or LS. The cohort includes individuals who had genetic testing, confirming either the presence or the absence of the familial pathogenic variant, and their untested relatives with unknown mutation status. Eligible participants may have had a cancer diagnosis, or they could be cancer-free at the time of enrolment in the study. Recruitment takes place at 8 different oncology and genetic testing centers in the German-, French-, and Italian-speaking regions of Switzerland. The study collects survey data designed to elicit factors that enhance cascade genetic testing and cancer surveillance for HBOC and LS. A subsample of participants has consented to provide narrative data regarding family communication of test results. For the purposes of this paper, we focused only on individuals who have had genetic testing for HBOC-associated pathogenic variants and accepted to provide narrative data.

Ethics Approval

The study protocol has been approved by the Ethics Committee of Northwest Switzerland (BASEC 2016-02052) and is publicly available (ClinicalTrials.gov NCT03124212) [30]. We also used available data from participants in the K-CASCADE study (ClinicalTrials.gov NCT04214210) in South Korea, which focuses on HBOC. K-CASCADE and the collaboration of the 2 studies has been approved by local ethics committees (Severance Hospital Institutional Review Board: 4-2020-0520). K-CASCADE is identical to the Swiss CASCADE in respect to scope, research design, participant eligibility criteria (except for age ≥19 years), and data collection methodology. Participants to K-CASCADE are recruited from 5 hospitals in South Korea [31].

Narrative Data

Narrative data included in this paper were collected from 44 individuals living in Switzerland and 9 in South Korea. The in-depth interviews were conducted between April 2019 and June 2021 either face to face or online (after April 2020 due to the COVID-19 pandemic) by trained research staff in German, French, Italian, English, and Korean using the same interview guide. Interview questions were designed to explore general communication patterns within family networks and specific experiences and barriers of family communication regarding genetic risk including discussions with health care providers. Examples of questions included in the interview guide are “What are some issues (barriers) that people might experience, related to sharing genetic risk information with family members?” and “Think of your own experience of (not) sharing genetic risk information with family members. What did you do and how did you decide about it?” Interviews were recorded, and all narrative data were transcribed verbatim in the original language in Microsoft Word and translated into English for this paper.

Survey Data

Survey data were collected on an ongoing basis, starting in fall 2017 and occurring approximately 18-24 months apart. Self-administered surveys assessed demographic and clinical characteristics [30]. The surveys also included investigator-developed items that have been associated with family communication and intention to inform relatives about genetic cancer risk. These items assess informational support among family members, preference for patient-mediated communication of genetic testing results, and perceived utility of genetic testing for relatives (Textbox 1). These items are scored on 7-point Likert-type scales ranging from 1 “Strongly Disagree” to 7 “Strongly Agree.” Respondents also completed the Informing Relatives Inventory (IRI), a 37-item scale assessing knowledge, motivation, and self-efficacy to disclose genetic cancer risk to relatives [32]. IRI items are also scored on a 7-point Likert-type scale, with higher overall score indicating greater intention to inform relatives about genetic cancer risk.

Textbox 1. Items from the CASCADE baseline survey used for this study.

Demographic characteristics

- Age
- Sex (female/male)
- Education level (elementary-, high school–, or technical school–graduate or academic degree)
- Marital status (married or living as married, single, divorced or separated, or widowed)
- Employment status (working full time, nonworking, or retired)

Clinical characteristics

- Cancer status (affected or never diagnosed with cancer)
- Genetic testing result (positive or negative for the familial pathogenic variant)

Family communication

- "In our family when I have a health problem there is great willingness to share information with each other"
- "I would prefer not to discuss about genetic testing results with anyone in my family"
- "If you have blood relatives, would it be useful for them to have genetic testing?"

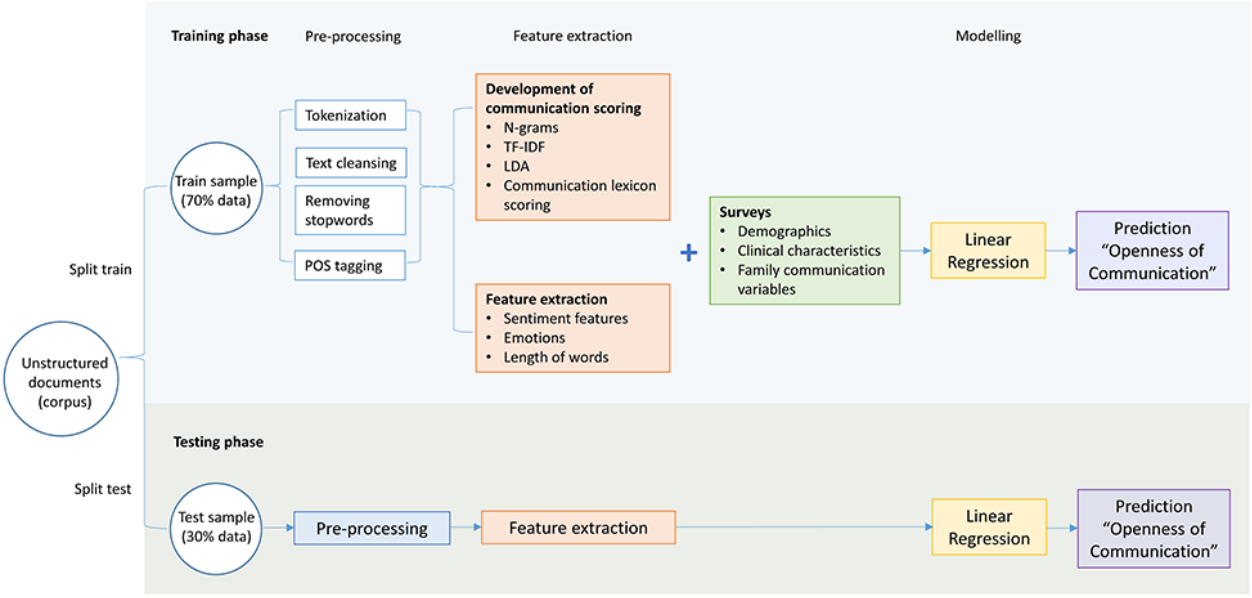
Data Analysis Overview

First, we examined narratives to assess "openness of communicating" genetic test results and cancer risk with relatives and with health care providers. Second, we categorized the text of each narrative as describing either a positive or negative sentiment toward experiences with genetic testing and health care services. Third, we examined whether demographic and clinical characteristics and sentiment, as expressed in the narrative, can predict "openness of communicating" genetic risk from tested individuals to relatives.

NLP Model Development

The ability of NLP to identify and predict different levels of "openness of communication" was evaluated following a multistep framework (Figure 1), which was divided into three phases: (1) preprocessing, (2) training, and (3) performance evaluation. All computations were performed in R software (version 3.6.3; R Foundation for Statistical Computing) [33]. We have made our analysis publicly available through the Zenodo open data repository [34].

Figure 1. Phases of developing the natural language processing (NLP) algorithm: (1) preprocessing, (2) training, and (3) performance evaluation. LDA: latent Dirichlet analysis; POS: part of speech; TF-IDF: term frequency–inverse document frequency.



Preprocessing Phase

To start data processing, we broke down each text into individual tokens. We then applied functions to remove stop words and special characters. All texts were converted to lower case. We also applied part-of-speech tagging to extract phrases from the text corpus, used a latent Dirichlet allocation model to generate the most appropriate topics, and computed the term frequency–inverse document frequency to indicate the significance of a word in the text corpus [35,36].

Creation of a Lexicon and a Score for "Openness of Communication"

To develop an "openness of communication" score, we built a lexicon containing words and phrases linked to communication (for example, "difficulties in communication" and "excellent communication") and classified them as positive or negative. After completing the preprocessing phase, we extracted N-grams from the text corpus. N-grams refer to single words (unigrams) or a combination of 2 or 3 words (bigrams or trigrams) associated with the outcome of interest, ie, "openness of communication." To further enrich the lexicon, we applied the same process in a US-based sample of 123 narratives related to experiences with HBOC genetic cancer risk. This database includes narrative data collected between January 2013 to September 2016 from women and men who are carriers of HBOC-associated variants [26]. The semistructured interviews inquired about experiences with genetic counseling, genetic testing, and family communication patterns. We enriched the lexicon with supplementary words related to communication identified in an online thesaurus [37]. The final lexicon we created contained 532 items (132 unigrams, 215 bigrams, and 185 trigrams). Two members of the research team independently created the scoring of N-grams in the lexicon as positive or negative without considering the context of the phrases in the interviews. Specifically, they evaluated each item on a 7-point scale on how favorable the items measure "openness of communication." Scoring values ranged from –3 (extremely strong negative word related to communication) to +3 (extremely strong positive word related to communication). In cases of disagreement, the final value was calculated by averaging the 2 values given by the 2 raters rounding to the greater nearest integer. The final "openness of communication" score assigned to the transcript of each narrative was developed by matching N-grams to the lexicon and summing up the corresponding scores. To ensure the robustness of the above scoring process, we calculated the Pearson correlation coefficient between the "openness of communication" scores we created with the IRI overall score. This correlation was examined only on Swiss data because Korean IRI scores were not available at the time of this analysis.

Sentiment Analysis and Attitude Toward Family Communication of Genetic Risk

To categorize the text of each narrative as describing either a positive or negative attitude toward genetic testing and health care services and to capture the overall emotional valence of the narrative, we used 3 common lexicons for text sentiment analysis: AFINN, Bing Liu, and the National Research Council Canada (NRC) Emotion Lexicon. The AFINN lexicon contains words with a score between –5 and +5, with negative and positive scores indicating negative and positive sentiments, respectively [38]. The Bing Liu lexicon classifies words into conveying a positive or a negative sentiment [1]. The NRC Emotion Lexicon estimates a sentiment score (positive and negative sentiment) based on 8 emotions. Positive emotions include anticipation, joy, surprise, and trust, whereas negative emotions include anger, disgust, fear, and sadness [39,40]. We also calculated an overall net sentiment expressed in each narrative, based on the difference between overall positive sentiment minus overall negative sentiment. An overall positive score meant that the individual expressed more positive sentiment in the narrative than negative, and vice versa.

Training Phase

For developing the model, the overall data set was split randomly, with 70% of data used in the training phase by using the "openness of communication" score as the dependent variable. To examine whether the demographic and clinical characteristics and sentiment features of each narrative predicted "openness of communication" scores, we used a linear regression model based on the following steps. Initially we performed a univariate analysis to identify those independent variables exhibiting more than 60% absolute correlation with one another. These variables were excluded to avoid multicollinearity. Then, we continued with a multivariate analysis using a stepwise linear regression to identify possible predictors of the dependent variable and remove nonsignificant independent variables. As an alternative model, we attempted to use an artificial neural network. We built a fully connected network with 1 hidden layer, 1 input and 1 output layer, and 5 neurons. Optimization was done through the Broyden-Fletcher-Goldfarb-Shanno method. Early stopping was utilized to avoid overfitting. However, we ended up discarding the artificial neural network from the analysis because it showed no improvement compared to the linear regression. Finally, the performance of the models was evaluated using the area under the curve (AUC), sensitivity, specificity, and root mean square error (RMSE).

Testing Phase

In this phase, we tested the model using the remaining 30% of the database (validation cohort). The performance of the models was evaluated using the same metrics as in the training phase, ie, AUC, sensitivity, specificity, and RMSE.

Results

Description of the Sample

Narrative and survey data from 53 individuals are included in this paper. Participants were aged 32-76 years. Most were female (47/53, 89%), married (41/53, 77%), and carriers of the familial pathogenic variant (51/53, 96%). Approximately 2 in 3 (32/53, 60%) had a prior diagnosis of cancer (Table 1). The Swiss and the Korean samples were not statistically different in respect to age (P=.71), prior cancer diagnosis (P=.38), educational level (P=.17), and employment status (P=.14).

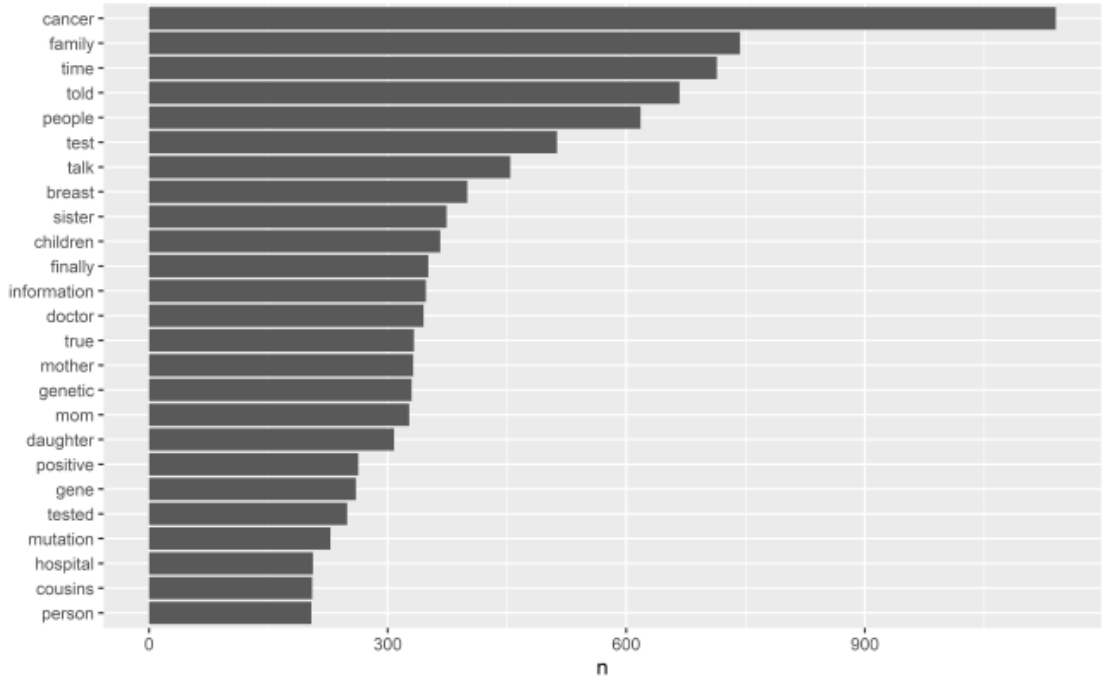
Table 1. Sociodemographic and clinical data of participants (N=53).

Characteristic	Value
Age (years), mean (SD)	53.3 (12.1)
Sex, n (%)	
Female	47 (89)
Education, n (%)	
Attended elementary/high school	5 (9)
High school graduate	14 (26)
Technical school graduate	13 (24)
University degree/postgraduate degree	21 (40)
Marital status, n (%)	
Married/living as married	41 (77)
Single	4 (8)
Divorced/separated/widowed	8 (15)
Employed full or part time (yes), n (%)	34 (64)
Cancer status, n (%)	
Previous cancer, one or more diagnoses	32 (60)
Never been diagnosed with cancer	21 (40)
Genetic test result, n (%)	
Positive for the familial pathogenic variant	51 (96)
Negative for the familial pathogenic variant	2 (4)

Description of the "Openness of Communication" Score and the Narrative Data

The average "openness of communication" score was 29.8 (SD 19.5; range –9 to 76), indicating an overall trend toward open communication. Narratives from these 53 individuals included 5837 unique unigrams, 4183 bigrams, and 654 trigrams. The most frequently appearing nontrivial words are shown in Figure 2. Based on the NRC Emotion Lexicon, the 10 most common positive words were "time," "true," "children," "talk," "finally," "information," "positive," "doctor," "understand," and "daughter". The 10 most common negative words were "cancer," "sick," "feel," "risk," "negative," "died," "difficult," "fear," "disease," and "bad."

Figure 2. The most frequent words identified in narratives.



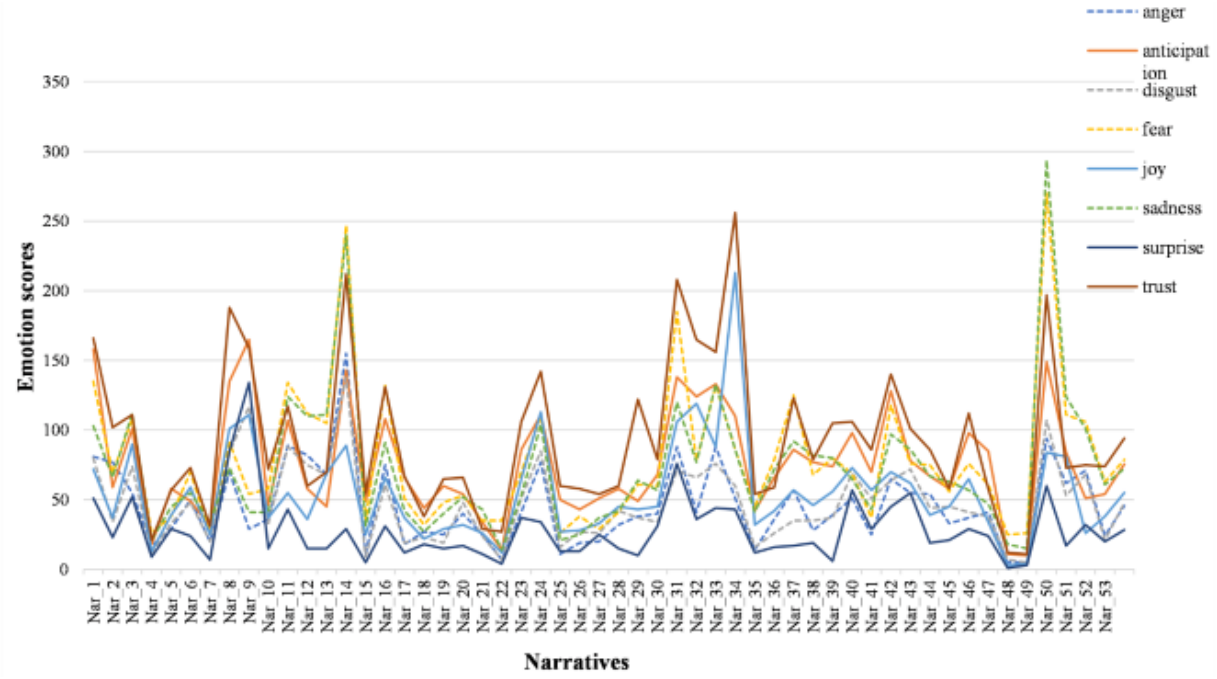
Validating the Relationship of "Openness of Communication" scores with IRI

The correlation coefficient between the "openness of communication" score and IRI in the Swiss data was r=0.46, indicating a moderate positive correlation.

Attitude Toward Genetic Testing and Health Care Services

Attitude toward genetic testing and health care services varied among participants, but it was overall positive. "Trust" appeared as the strongest positive emotion, whereas "fear" and "sadness" appeared as the strongest negative emotions in the text corpus based on the NRC Emotion Lexicon. The least perceived emotions were "surprise" and "anger." Figure 3 describes the frequencies of words identified in the corpus for each emotion.

Figure 3. Frequencies of words identified for each emotion (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust) based on the National Research Council Canada Emotion Lexicon.



Prediction of "Openness of Communication" Score

The R^2 for the overall model was 0.87 (adjusted R^2=0.85; P<.001). A stepwise linear regression identified 5 significant predictors of "openness of communication" score, ie, the overall net sentiment of the narrative and fear, which were obtained based on the NRC Emotion Lexicon; informational support among family members; educational level; and being single (Table 2). Specifically, findings showed that both the higher overall net sentiment score of the narrative (P=.007) and also greater fear (P=1.97 × 10^{-5}) were strongly associated with higher "openness of communication" scores. There was a positive correlation between "openness of communication" score and the statement "In our family when I have a health problem there is great willingness to share information with each other" (P=.005). Participants with nonacademic education were also more likely to communicate genetic risk with their relatives (P=.02). Lastly, there was a positive correlation between being single and "openness of communication" scores (P=.047).

Table 2. Results of the linear regression analysis predicting "openness of communication."

Variables	Estimate	SE	*t* test[a] (*df*)	*P* value
Being single	19.782	9.574	2.066 (1)	.047[b]
Academic education	–10.387	4.256	–2.44 (1)	.02[b]
Fear	0.204	0.041	4.954 (1)	1.97×10^{-5}[c]
Informational support	11.392	3.790	3.006 (1)	.005[d]
Net sentiment score of the narrative	0.260	0.091	2.861 (1)	.007[d]

[a]2-tailed *t* test.

[b]Significance level: *P*<.05.

[c]Significance level: *P*<.001.

[d]Significance level: *P*<.01.

Model Performance

The predictive accuracy of the model using a stepwise linear regression for the training and testing data sets reached 0.85 (AUC=0.92, specificity=0.86, and sensitivity=0.82) and 0.72 (AUC=0.69, specificity=0.62, and sensitivity=0.83), respectively. Figure 4 presents the receiver operating characteristic curves that visualize the accuracy improvement between the training and testing data sets applying linear regression.

The predicted values are plotted against the target values and are shown on a scatter plot for the linear regression model (Figure 5). The linear regression model achieved RMSEs of 11.76 (training data set) and 16.04 (testing data set). In this case, our model performs more accurately when it yields lower values of RSME.

Figure 4. Receiver operating characteristic (ROC) curves of the training (A) and testing (B) model predicting "openness of communication" applying linear regression. AUC: area under the curve.

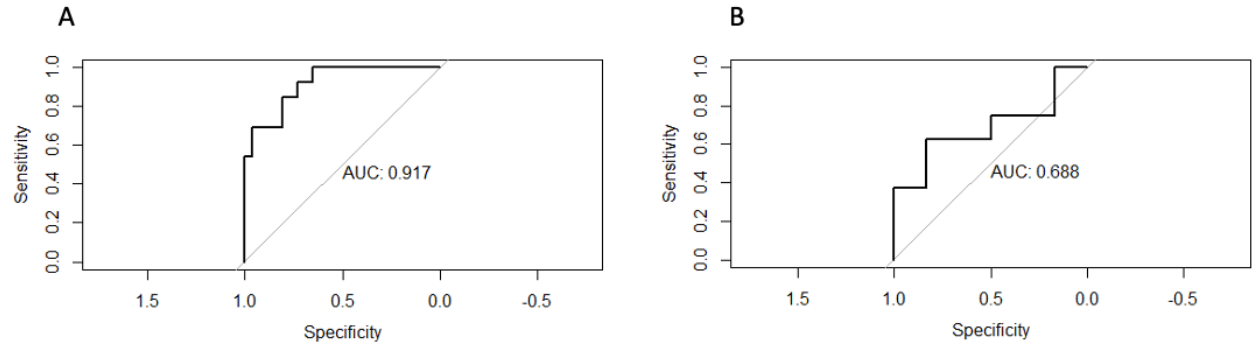
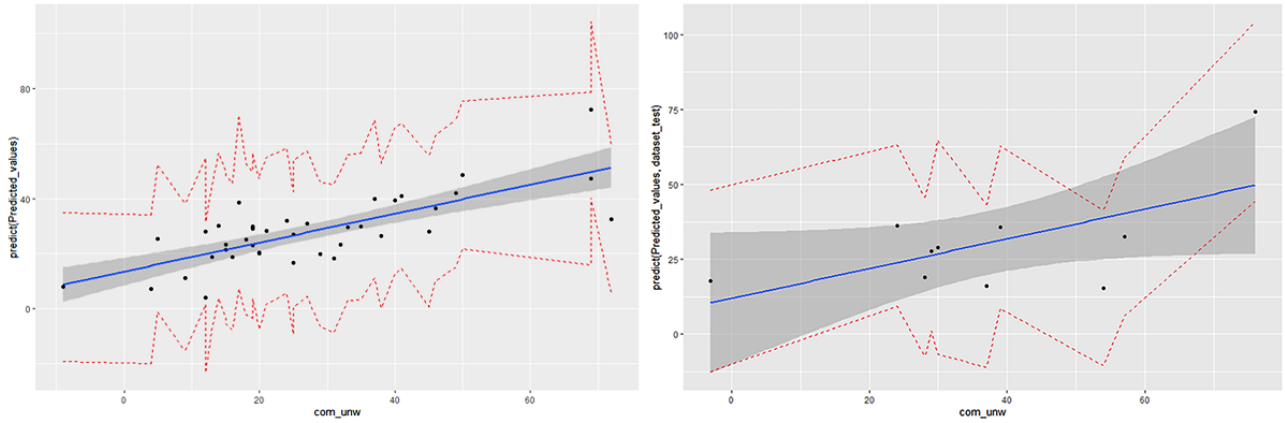


Figure 5. Scatter plot of predicted values against target values with 95% confidence and prediction intervals for the training (A) and testing (B) data sets applying linear regression.



Discussion

Principal Findings

We analyzed 53 narratives regarding intrafamilial communication of genetic cancer risk associated with HBOC. NLP enabled the analysis of unstructured narratives from different languages and identified the most frequently used words or combination of words describing openness in family communication of genetic cancer risk. This was the first study in which we applied NLP and sentiment analysis to better understand factors driving open intrafamilial communication regarding genetic cancer risk. Our findings showed that sentiment plays a crucial role and that emotions are a pervasive feature that predict intrafamilial communication in this particular population. Sentiment analysis performed on all interviews provided scores demonstrating positive or negative emotional valence, which were highly predictive of the direction of intrafamilial communication in this context. The higher overall net sentiment scores predicted greater openness in intrafamilial communication, whereas the lower overall net sentiment scores predicted closed or absent communication. This finding provides insights consistent with social penetration theory related to self-disclosure of carrying a cancer-causing genetic variant [41,42]. The depth of self-disclosure, ie, the degree to which the individual reveals personal and private information involving unusual traits and painful memories, reflects the degree of intimacy of a relationship. In the context of HBOC intrafamilial communication, self-disclosure of personal genetic information may be opposed by the desire to retain privacy and to avoid creating uncertainty and unpredictability in interpersonal relationships. Anticipating future negative emotions, such as regret or conflict, categorizes genetic risk information as a considerable emotional threat [43]. This finding was captured in our analysis as the overall net sentiment of each narrative, and its predictive value was confirmed based on the performance of our models. Taken together, findings indicate that sentiment can be used to frame genetic cancer risk as an opportunity for proactive risk reduction and for enhancing technology-mediated HBOC intrafamilial communication.

Our linear regression model explained more than 80% of the variance in openness of communication and achieved good performance in both the training and the testing samples. Our findings show that NLP was highly accurate in analyzing unstructured narratives from individuals of different cultural and linguistic backgrounds (Swiss German, French, Italian, English, and Korean) and in quantifying openness of communication in intrafamilial discussions about genetic cancer risk. The "openness of communication" scores were also validated against IRI. IRI was developed on the premise that increased genetic knowledge, positive motivation, and increased self-efficacy are prerequisites of increased intention to inform relatives about genetic risk. Although "intention to inform relatives" is closely related to "openness of communicating genetic risk," the 2 concepts are not identical, which was also confirmed in our data with a moderate positive correlation between the 2 scores. An individual may have high intention to inform relatives about their genetic risk despite difficulties in communication within their family.

Creating a new lexicon for openness in communication enriched with terms from different sources contributes to the innovation of our approach and the generalizability and applicability of our findings. Our lexicon can be further used and expanded in future projects, providing a solid foundation for the use of NLP in the growing field of research in interpersonal communication, focusing on family communication and health care and technology-mediated communication [44]. Sentiment analysis can be further utilized in the era of precision medicine and precision public health for message tailoring and message framing. Extracting sentiment polarities can be highly informative in improving consumer experiences when using digital health platforms in promoting precision public health campaigns. For example, trust in the health care system has been associated with use of cancer surveillance, whereas conflicting messages from providers create a sense of disorientation and mistrust [45-48].

Our findings also indicated a greater likelihood of open intrafamilial communication in those who were single, had a nonacademic education, and higher informational support within their family network. These findings should be interpreted with caution and should be replicated with analyses of narratives from larger, and possibly more diverse, samples.

Strengths and Limitations

Studies in different domains have also considered sentiment for analyzing textual communication in social media such as Twitter or Facebook [5,9,11]. However, one significant strength of our approach was that narrative data were combined with the demographic and clinical characteristics of participants, which can increase the applicability of findings. Another important strength was the use of several sentiment lexicons to select the most suitable for this context. Sentiment scores originating from the NRC Emotion Lexicon were the most appropriate to predict "openness of communication," whereas the other 2 sentiment lexicons (AFINN and Bing Liu) were highly correlated, resulting in a predictive algorithm of lesser importance. Studies have shown that the selection of inappropriate lexicons may impact prediction performance [39,40]. Finally, NLP can automate parts of text analysis and can be used as an assisting tool to help researchers navigate through large volumes of text data.

One limitation of our study was the small sample size and the size of the available corpus, which did not allow us to include possible significant covariates and to fully explore the potential of the NLP methodology, including sentence structure and length of words. Despite this limitation, the results of our study can be used as indicators of various narrative features, such as overall sentiment and fear, which can be important predictors of interpersonal communication and self-disclosure in this specific population. Important features of NLP analysis, such as sentence structure and length of words, can be investigated with a larger number of narratives and larger number of corpora. The analytical approach we describe in this paper can be further improved by using larger samples. Further development of a robust model will advance a more precise assessment and reach higher accuracy.

Conclusions

We demonstrate how various features from narratives can be used to predict "openness of communication" in individuals carrying a pathogenic variant connected to HBOC. Although our methodology requires further exploration and our findings require replication with larger samples, this is an important first step to understand how individuals and the public may react in discourses involving communication of genetic cancer risk. Overall, this experimental analysis provides evidence that our approach is promising and can be further used in the field of technology-mediated communication and precision public health.

Acknowledgments

We would like to thank Monica Aceti, PhD; Andrea Kaiser-Grolimund, PhD; and Carla Pedrazzani, MSc; and the recruitment team of K-CASCADE and the CASCADE Consortia for their valuable help regarding data collection and ongoing recruitment. This research was supported by the Swiss Cancer League (KLS-4294-08-2017); Swiss Cancer Research Foundation (KFS-5293-02-2021); DIALOGUE (IZKSZ3_188408/1); and the University of Basel, Office of the Vice Rector of Research (2016) to MCK.

The members of the CASCADE Consortium are Monica Aceti, Souria Aissaoui, Nicole Bürki, Pierre O Chappuis, Muriel Fluri, Rossella Graffeo-Galbiati, Karl Heinimann, Viola Heinzelmann-Schwarz, Helen Koechlin, Christian Kurzeder, Ashley Machen, Christian Monnerat, Carla Pedrazzani, Nicole Probst-Hensch, Manuela Rabaglio, Simon Wieser and Ursina Zürrer-Härdi.

Authors' Contributions

VB, RCGP, GL, and MCK contributed to conceptualization. VB, RCGP, RS, and MCK contributed to methodology. VB contributed to formal analysis and visualization. VB and MCK contributed to writing—original draft preparation. VB, RCGP, SK, RS, SHB, MCZ, GL, FC, and MCK contributed to writing—review and editing. MCK contributed to supervision. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

References

1. Hu M, Liu B. Mining and summarizing customer reviews. 2004 Aug 22 Presented at: KDD '04: tenth ACM SIGKDD international conference on Knowledge discovery and data mining; August 22-25, 2004; Seattle, WA p. 168-177. [doi: 10.1145/1014052.1014073]
2. Pang B, Lee L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2008 Jul 07;2(1–2):1-135. [doi: 10.1561/1500000011]
3. Solangi YA, Solangi ZA, Aarain S, Abro A, Mallah GA, Shah A. Review on natural language processing (NLP) and its toolkits for opinion mining and sentiment analysis. 2018 Presented at: 2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS); November 22-23, 2018; Bangkok, Thailand p. 1-4. [doi: 10.1109/icetas.2018.8629198]
4. Kastrati Z, Dalipi F, Imran AS, Pireva Nuci K, Wani MA. Sentiment analysis of students' feedback with NLP and deep learning: a systematic mapping study. Appl Sci 2021 Apr 28;11(9):3986. [doi: 10.3390/app11093986]
5. Kaur H, Ahsaan SU, Alankar B, Chang V. A proposed sentiment analysis deep learning algorithm for analyzing COVID-19 tweets. Inf Syst Front 2021 Apr 20;23(6):1417-1429 [FREE Full text] [doi: 10.1007/s10796-021-10135-7] [Medline: 33897274]
6. Manguri KH, Ramadhan RN, Mohammed Amin PR. Twitter sentiment analysis on worldwide COVID-19 outbreaks. Kurdistan Journal of Applied Research 2020 May 19;5(3):54-65. [doi: 10.24017/covid.8]
7. Mishev K, Gjorgjevikj A, Stojanov R, Mishkovski I, Vodenska I, Chitkushev L. Performance evaluation of word and sentence embeddings for finance headlines sentiment analysis. 2019 Oct 14 Presented at: Performance evaluation of word and sentence embeddings for finance headlines sentiment analysis. International Conference on ICT Innovations; : Springer; October 17-19, 2019; Ohrid, North Macedonia p. 161-172. [doi: 10.1007/978-3-030-33110-8_14]
8. Petropoulos A, Siakoulis V. Can central bank speeches predict financial market turbulence? evidence from an adaptive NLP sentiment index analysis using XGBoost machine learning technique. Central Bank Review 2021 Dec;21(4):141-153. [doi: 10.1016/j.cbrev.2021.12.002]
9. Paltoglou G, Thelwall M. Twitter, MySpace, Digg: unsupervised sentiment analysis in social media. ACM Trans Intell Syst Technol 2012 Sep;3(4):1-19. [doi: 10.1145/2337542.2337551]
10. Bhaskar J, Sruthi K, Nedungadi P. Enhanced sentiment analysis of informal textual communication in social media by considering objective words and intensifiers. 2014 Presented at: nternational Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014); May 9-11, 2014; Jaipur, India p. 1-6. [doi: 10.1109/icraie.2014.6909220]
11. Sailunaz K, Alhajj R. Emotion and sentiment analysis from Twitter text. J Comput Sci 2019 Sep;36:101003. [doi: 10.1016/j.jocs.2019.05.009]

12. Liu X, Jiang S, Sun M, Chi X. Examining patterns of information exchange and social support in a web-based health community: exponential random graph models. J Med Internet Res 2020 Sep 29;22(9):e18062 [FREE Full text] [doi: 10.2196/18062] [Medline: 32990628]
13. Lu Y, Wu Y, Liu J, Li J, Zhang P. Understanding health care social media use from different stakeholder perspectives: a content analysis of an online health community. J Med Internet Res 2017 Apr 07;19(4):e109 [FREE Full text] [doi: 10.2196/jmir.7087] [Medline: 28389418]
14. Cabling ML, Turner JW, Hurtado-de-Mendoza A, Zhang Y, Jiang X, Drago F, et al. Sentiment analysis of an online breast cancer support group: communicating about Tamoxifen. Health Commun 2018 Sep;33(9):1158-1165 [FREE Full text] [doi: 10.1080/10410236.2017.1339370] [Medline: 28678549]
15. Bian J, Zhao Y, Salloum RG, Guo Y, Wang M, Prosperi M, et al. Using social media data to understand the impact of promotional information on laypeople's discussions: a case study of Lynch syndrome. J Med Internet Res 2017 Dec 13;19(12):e414 [FREE Full text] [doi: 10.2196/jmir.9266] [Medline: 29237586]
16. Hyun S, Johnson SB, Bakken S. Exploring the ability of natural language processing to extract data from nursing narratives. Comput Inform Nurs 2009;27(4):215-23; quiz 224 [FREE Full text] [doi: 10.1097/NCN.0b013e3181a91b58] [Medline: 19574746]
17. Cammel SA, de Vos MS, van Soest D, Hettne KM, Boer F, Steyerberg EW, et al. How to automatically turn patient experience free-text responses into actionable insights: a natural language programming (NLP) approach. BMC Med Inform Decis Mak 2020 May 27;20(1):97 [FREE Full text] [doi: 10.1186/s12911-020-1104-5] [Medline: 32460734]
18. Vehviläinen-Julkunen K, Turpeinen S, Kvist T, Ryden-Kortelainen M, Nelimarkka S, Enshaeifar S, INEXCA Consortium. Experience of ambulatory cancer care: understanding patients' perspectives of quality using sentiment analysis. Cancer Nurs 2021;44(6):E331-E338. [doi: 10.1097/NCC.0000000000000845] [Medline: 32618620]
19. Cox DJ, Garcia-Romeu A, Johnson MW. Predicting changes in substance use following psychedelic experiences: natural language processing of psychedelic session narratives. Am J Drug Alcohol Abuse 2021 Jul 04;47(4):444-454. [doi: 10.1080/00952990.2021.1910830] [Medline: 34096403]
20. Badal VD, Graham SA, Depp CA, Shinkawa K, Yamada Y, Palinkas LA, et al. Prediction of loneliness in older adults using natural language processing: exploring sex differences in speech. Am J Geriatr Psychiatry 2021 Aug;29(8):853-866 [FREE Full text] [doi: 10.1016/j.jagp.2020.09.009] [Medline: 33039266]
21. Schwiter R, Rahm AK, Williams JL, Sturm AC. How can we reach at-risk relatives? efforts to enhance communication and cascade testing uptake: a mini-review. Curr Genet Med Rep 2018 Apr 19;6(2):21-27. [doi: 10.1007/s40142-018-0134-0]
22. Ignatius E, Kokkonen M. Factors contributing to verbal self-disclosure. Nord Psychol 2012 Jul 11;59(4):362-391. [doi: 10.1027/1901-2276.59.4.362]
23. Baxter LA, Braithwaite DO. Relational dialectics theory: crafting meaning from competing discourses. In: Engaging Theories in Interpersonal Communication: Multiple Perspectives. Thousand Oaks, CA: SAGE Publications; 2008:01-61.
24. Braithwaite DO. Engaging Theories in Interpersonal Communication: Multiple Perspectives. Thousand Oaks, CA: SAGE Publications; 2005:349-362.
25. Dean M, Tezak AL, Johnson S, Pierce JK, Weidner A, Clouse K, et al. Sharing genetic test results with family members of BRCA, PALB2, CHEK2, and ATM carriers. Patient Educ Couns 2021 Apr;104(4):720-725 [FREE Full text] [doi: 10.1016/j.pec.2020.12.019] [Medline: 33455826]
26. Dwyer AA, Hesse-Biber S, Flynn B, Remick S. Parent of origin effects on family communication of risk in BRCA+ women: a qualitative investigation of human factors in cascade screening. Cancers (Basel) 2020 Aug 17;12(8):2316 [FREE Full text] [doi: 10.3390/cancers12082316] [Medline: 32824510]
27. Mahdavi M, Nassiri M, Kooshyar MM, Vakili-Azghandi M, Avan A, Sandry R, et al. Hereditary breast cancer; genetic penetrance and current status with BRCA. J Cell Physiol 2019 May 14;234(5):5741-5750. [doi: 10.1002/jcp.27464] [Medline: 30552672]
28. Chivers Seymour K, Addington-Hall J, Lucassen A, Foster C. What facilitates or impedes family communication following genetic testing for cancer risk? A systematic review and meta-synthesis of primary qualitative research. J Genet Couns 2010 Aug;19(4):330-342 [FREE Full text] [doi: 10.1007/s10897-010-9296-y] [Medline: 20379768]
29. Srinivasan S, Won NY, Dotson WD, Wright ST, Roberts MC. Barriers and facilitators for cascade testing in genetic conditions: a systematic review. Eur J Hum Genet 2020 Dec;28(12):1631-1644 [FREE Full text] [doi: 10.1038/s41431-020-00725-5] [Medline: 32948847]
30. Katapodi MC, Viassolo V, Caiata-Zufferey M, Nikolaidis C, Bührer-Landolt R, Buerki N, et al. Cancer predisposition cascade screening for hereditary breast/ovarian cancer and Lynch syndromes in Switzerland: study protocol. JMIR Res Protoc 2017 Sep 20;6(9):e184 [FREE Full text] [doi: 10.2196/resprot.8138] [Medline: 28931501]
31. Kim S, Aceti M, Baroutsou V, Bürki N, Caiata-Zufferey M, Cattaneo M, et al. Using a tailored digital health intervention for family communication and cascade genetic testing in Swiss and Korean families with hereditary breast and ovarian cancer: protocol for the DIALOGUE study. JMIR Res Protoc 2021 Jun 11;10(6):e26264 [FREE Full text] [doi: 10.2196/26264] [Medline: 34114954]
32. de Geus E, Aalfs CM, Menko FH, Sijmons RH, Verdam MGE, de Haes HCJM, et al. Development of the Informing Relatives Inventory (IRI): assessing index patients' knowledge, motivation and self-efficacy regarding the disclosure of

hereditary cancer risk information to relatives. Int J Behav Med 2015 Aug;22(4):551-560. [doi: 10.1007/s12529-014-9455-x] [Medline: 25515913]
33. R Core Team. R: a language and environment for statistical computing. The R Project for Statistical Computing. URL: https://www.r-project.org/ [accessed 2022-12-22]
34. Baroutsou V. Analysis code for the paper "predicting openness of communication in families with hereditary breast and ovarian cancer syndrome: natural language processing analysis". Zenodo. 2022 May 04. URL: https://doi.org/10.5281/zenodo.6517726 [accessed 2022-12-22]
35. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003 Mar 01;3:993-1022 [FREE Full text]
36. Aizawa A. An information-theoretic perspective of tf–idf measures. Information Processing & Management 2003 Jan;39(1):45-65. [doi: 10.1016/S0306-4573(02)00021-3]
37. Power Thesaurus. URL: https://www.powerthesaurus.org/ [accessed 2021-12-09]
38. Nielsen FÅ. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. arXiv Preprint posted online on March 15, 2011. [doi: 10.48550/arXiv.1103.2903]
39. Mohammad SM, Turney PD. Crowdsourcing a word-emotion association lexicon. Comput Intell 2013 Aug;29(3):436-465. [doi: 10.1111/j.1467-8640.2012.00460.x]
40. Mohammad SM, Turney PD. Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon. 2010 Jun Presented at: NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text; June 5, 2010; Los Angeles, CA p. 26-34.
41. Carpenter A, Greene K. Social penetration theory. In: The International Encyclopedia of Interpersonal Communication. New York, NY: John Wiley & Sons; Dec 01, 2015.
42. Masaviru M. Self-disclosure: theories and model review. Journal of Culture, Society and Development 2016;18:43-47 [FREE Full text]
43. Wöhlke S, Schaper M, Schicktanz S. How uncertainty influences lay people's attitudes and risk perceptions concerning predictive genetic testing and risk communication. Front Genet 2019 Apr 26;10:380 [FREE Full text] [doi: 10.3389/fgene.2019.00380] [Medline: 31080458]
44. Caughlin J, Koerner AF, Schrodt P, Fitzpatrick MA. Interpersonal communication in family relationships. In: Knapp ML, Daly JA, editors. The Sage Handbook of Interpersonal Communication. 4th ed. Thousand Oaks, CA: SAGE Publications; 2011:679-714.
45. Caiata-Zufferey M, Pagani O, Cina V, Membrez V, Taborelli M, Unger S, et al. Challenges in managing genetic cancer risk: a long-term qualitative study of unaffected women carrying BRCA1/BRCA2 mutations. Genet Med 2015 Sep;17(9):726-732 [FREE Full text] [doi: 10.1038/gim.2014.183] [Medline: 25503500]
46. Kaiser K, Rauscher GH, Jacobs EA, Strenski TA, Ferrans CE, Warnecke RB. The import of trust in regular providers to trust in cancer physicians among White, African American, and Hispanic breast cancer patients. J Gen Intern Med 2011 Jan;26(1):51-57 [FREE Full text] [doi: 10.1007/s11606-010-1489-4] [Medline: 20811783]
47. Katapodi MC, Pierce PF, Facione NC. Distrust, predisposition to use health services and breast cancer screening: results from a multicultural community-based survey. Int J Nurs Stud 2010 Aug;47(8):975-983. [doi: 10.1016/j.ijnurstu.2009.12.014] [Medline: 20089252]
48. Kenen RH, Shapiro PJ, Friedman S, Coyne JC. Peer-support in coping with medical uncertainty: discussion of oophorectomy and hormone replacement therapy on a web-based message board. Psychooncology 2007 Aug;16(8):763-771. [doi: 10.1002/pon.1152] [Medline: 17230435]

Abbreviations

AUC: area under the curve
HBOC: hereditary breast and ovarian cancer
IRI: Informing Relatives Inventory
LS: Lynch syndrome
NLP: natural language processing
NRC: National Research Council Canada
RMSE: root mean square error

Edited by A Mavragani; submitted 31.03.22; peer-reviewed by R Pozzar, ER Khalilian, R Zhao; comments to author 23.06.22; revised version received 11.07.22; accepted 24.11.22; published 19.01.23

Please cite as:
Baroutsou V, Cerqueira Gonzalez Pena R, Schweighoffer R, Caiata-Zufferey M, Kim S, Hesse-Biber S, Ciorba FM, Lauer G, Katapodi M, CASCADE Consortium
Predicting Openness of Communication in Families With Hereditary Breast and Ovarian Cancer Syndrome: Natural Language Processing Analysis
JMIR Form Res 2023;7:e38399
URL: https://formative.jmir.org/2023/1/e38399
doi: 10.2196/38399
PMID: