Data Management and Sharing Plan

Element 1: Data Type:

A. Types and amount of scientific data expected to be generated in the project:

Summarize the types and estimated amount of scientific data expected to be generated in the project.

Social media posts and comments: Approximately 300,000 publicly available social media posts (including associated images and videos) and their comments, obtained via a third-party data mining vendor. These will be processed, de-identified, and coded for various attributes (e.g., topic, sentiment, accuracy, use of personal narrative).

Survey data: Quantitative survey responses from 500 young adult participants (ages 18-34) enrolled in an Instagram-based intervention. The survey will capture demographic information, independent and dependent variable measures, and intervention group assignments. Data will be recorded as necessary to address missing values and other processing needs.

User analytics: Program monitoring data from 500 participants, including metrics such as views, comments, shares, and engagement patterns within intervention and control groups. Coded results of major themes and post characteristics will also be included. All identifiers will be removed. The anticipated total data volume is estimated at approximately 50-100 GB, primarily due to the large volume of social media content (including multimedia files).

B. Scientific data that will be preserved and shared, and the rationale for doing so:

Describe which scientific data from the project will be preserved and shared and provide the rationale for this decision.

All de-identified and processed scientific data will be preserved and shared, including:

- The coded, de-identified dataset of social media posts and comments (with associated categorical variables and annotations).
- The de-identified participant survey dataset, including group assignments and recorded variables.
- The de-identified user analytics dataset, including engagement metrics and coded qualitative themes.
- The rationale for sharing these data is to facilitate reproducibility, secondary analyses, and advancement of knowledge in cancer prevention and health communication research.
- Raw social media posts with potentially identifying information will not be shared; only de-identified, coded datasets will be disseminated in accordance with ethical and legal guidelines.

C. Metadata, other relevant data, and associated documentation:

Briefly list the metadata, other relevant data, and any associated documentation (e.g., study protocols and data collection instruments) that will be made accessible to facilitate interpretation of the scientific data.

The following documentation will be shared to facilitate data interpretation:

- Data dictionaries for all shared datasets, describing variables, coding schemes, and value labels.
- Study protocol and survey instruments.
- Codebooks for thematic coding and sentiment analysis of social media data.
- Documentation of data processing and de-identification procedures.
- README files describing dataset structure, file formats, and any analytic code provided.

Element 2: Related Tools, Software and/or Code:

State whether specialized tools, software, and/or code are needed to access or manipulate shared scientific data, and if so, provide the name(s) of the needed tool(s) and software and specify how they can be accessed.

The shared datasets will be provided in standard, non-proprietary formats (e.g., CSV, TXT, and JSON for text and coded data; MP4 or JPEG for example media files, where appropriate and permitted). No specialized software is required to access the primary datasets. If analytic code (e.g., for data cleaning or sentiment analysis) is shared, it will be provided as commented R or Python scripts, which are open source and freely available. Instructions for code use and software installation will be included in the documentation.

Element 3: Standards:

State what common data standards will be applied to the scientific data and associated metadata to enable interoperability of datasets and resources and provide the name(s) of the data standards that will be applied and describe how these data standards will be applied to the scientific data generated by the research proposed in this project. If applicable, indicate that no consensus standards exist.

The project will apply the following data and metadata standards to promote interoperability: - Variable naming and coding will adhere to established standards where available (e.g., NIH Common Data Elements for demographic variables). - Survey data will be structured according to the Data Documentation Initiative (DDI) standard. - Social media and user analytics data will be formatted following best practices for social media research data (e.g., including TIMESTAMP, POST_ID, CODED_VARIABLES, etc.). - Metadata will be provided in accordance with the FAIR principles (Findable, Accessible, Interoperable, Reusable). Where consensus standards do not exist (e.g., for certain qualitative codes), detailed codebooks and documentation will be provided.

Element 4: Data Preservation, Access, and Associated Timelines:

A. Repository where scientific data and metadata will be archived:

Provide the name of the repository(ies) where scientific data and metadata arising from the project will be archived.

The scientific data and metadata will be archived in the NCI-supported Cancer Data Access System (CDAS) and/or the NIMH Data Archive (NDA), both of which are NIH-approved repositories for human subjects and behavioral data. Supplementary materials (e.g., analytic code) will also be deposited on the Open Science Framework (OSF) or Zenodo for broader accessibility.

***B.* How scientific data will be findable and identifiable:**

Describe how the scientific data will be findable and identifiable, i.e., via a persistent unique identifier or other standard indexing tools.

Data will be assigned persistent unique identifiers (e.g., Digital Object Identifiers [DOI]) by the selected repositories. Metadata records will be indexed and searchable through the repository platforms and referenced in associated publications.

***C.* When and how long the scientific data will be made available:**

Describe when the scientific data will be made available to other users (i.e., no later than the time of an associated publication or end of the performance period, whichever comes first) and for how long data will be available.

Scientific data will be made available no later than the time of publication of the primary results or at the end of the NIH project performance period, whichever comes first. The data will remain available for a minimum of 10 years, in accordance with NIH and repository policies.

Element 5: Access, Distribution, or Reuse Considerations

A. Factors affecting subsequent access, distribution, or reuse of scientific data:

NIH expects that in drafting Plans, researchers maximize the appropriate sharing of scientific data. Describe and justify any applicable factors or data use limitations affecting subsequent access, distribution, or reuse of scientific data related to informed consent, privacy and confidentiality protections, and any other considerations that may limit the extent of data sharing.

Data sharing will comply with informed consent provisions obtained from participants, which permit broad data sharing. All datasets will be de-identified to protect participant privacy. Restrictions may apply to raw multimedia files or social media content that cannot be fully de-identified or where redistribution is limited by platform terms of service. In such cases, only coded, de-identified data will be shared. Data use agreements may require users to commit to ethical use and not attempt re-identification.

B. Whether access to scientific data will be controlled:

State whether access to the scientific data will be controlled (i.e., made available by a data repository only after approval).

Access to de-identified survey and user analytics data will be controlled via the repository's data access committees to ensure compliance with consent and ethical guidelines. Coded social media datasets will also be available through controlled access, requiring data use agreements and approval. Metadata and documentation will be publicly accessible.

C. Protections for privacy, rights, and confidentiality of human research participants:

If generating scientific data derived from humans, describe how the privacy, rights, and confidentiality of human research participants will be protected (e.g., through de-identification, Certificates of Confidentiality, and other protective measures).

All datasets will be thoroughly de-identified prior to sharing, with direct and indirect identifiers removed or masked. Data will be reviewed for re-identification risk. The study will obtain a Certificate of Confidentiality from NIH. Only data for which broad sharing is permitted by participant consent will be shared. Requests for access will be reviewed by data access committees, and data use agreements will prohibit re-identification attempts.

Element 6: Oversight of Data Management and Sharing:

Describe how compliance with this Plan will be monitored and managed, frequency of oversight, and by whom at your institution (e.g., titles, roles).

Compliance with this Data Management and Sharing Plan will be overseen by the study's Principal Investigator, with support from the project's Data Steward and the Institutional Data Management Committee. Oversight activities will include quarterly reviews of data management practices, de-identification protocols, and sharing timelines. The Office of Research Compliance at the applicant institution will conduct annual audits of data sharing compliance. Progress will be reported to the NIH program officer as part of routine progress reports.