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.

The proposed study will generate several types of scientific data, including:

- Approximately 300,000 publicly available social media posts (including images and videos) with associated comments.
- Quantitative survey data from 500 young adult participants (age 18-34), which includes demographic characteristics and measures for all independent and dependent variables.
- User analytics such as views, comments, and shares from the intervention delivered via Instagram. The estimated total amount of data is expected to be substantial due to the large number of social media posts and participant responses.

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 collected data (social media posts, survey data, user analytics) will be preserved and shared because they are crucial for understanding the impact of the intervention on young adults' behaviors and attitudes towards cancer prevention. Sharing these data will facilitate replication studies, secondary analyses, and the development of new interventions.

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.

Metadata to be made accessible includes:

- Study protocols
- Data collection instruments (survey questions)
- Coding schemes for social media posts and comments (e.g., topic, sentiment, accuracy, use of personal narrative)
- Documentation on data processing and analysis methods
- Participant demographic information (de-identified)

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.

Specialized tools and software needed to access or manipulate shared scientific data include: - Social media data mining vendor's software for initial data collection - Statistical analysis software (e.g., R, Python) for data processing and analysis - Qualitative data analysis software (e.g., NVivo, Atlas.ti) for coding and analyzing social media posts and comments These tools can be accessed through academic licenses or open-source repositories.

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.

Common data standards to be applied include: - Data dictionary standards for survey data - Standardized coding schemes for qualitative data analysis (e.g., based on existing literature) - Compliance with FAIR principles (Findability, Accessibility, Interoperability, Reusability) for all shared data and metadata These standards will enable interoperability of datasets and resources, facilitating future research and meta-analyses.

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 National Cancer Institute's (NCI) Genomic Data Commons (GDC) or other NIH-approved repositories will be used to archive the scientific data and metadata arising from this project.

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.

Scientific data will be made findable and identifiable through:

- Assignment of a Digital Object Identifier (DOI)
- Use of standardized metadata formats
- Indexing in public databases and search engines

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.

The scientific data will be made available no later than the time of publication of the main findings or within 12 months after the end of the performance period, whichever comes first. Data will remain accessible for at least 10 years following the completion of the project to allow for long-term reuse and analysis.

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.

Given that participants have provided consent for broad data sharing, there are minimal limitations on access, distribution, or reuse of the scientific data. However, all shared data will be de-identified to protect participant privacy and confidentiality.

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 the scientific data will not be controlled beyond standard repository access protocols. Data will be made openly available to maximize sharing and reuse.

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).

To protect participant privacy, rights, and confidentiality:

- All identifiable information will be removed from shared data
- Participants' consent includes agreement with data sharing practices
- Certificates of Confidentiality will be obtained where applicable to further protect sensitive information

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 Plan will be monitored and managed by the Principal Investigator (PI) in collaboration with the institution's research administration office. Regular reviews (every 6 months) will ensure that data management and sharing practices adhere to this plan and all relevant regulations and policies.