Plan Overview

A Data Management Plan created using DMPTool

DMP ID: https://doi.org/10.48321/D1BK5T

Title: Using natural language processing to determine predictors of healthy diet and physical activity behavior change in ovarian cancer survivors

Creator: Damian Yukio Romero Diaz - **ORCID:** 0000-0003-4661-0296

Affiliation: University of Arizona (arizona.edu)

Principal Investigator: Steven Bethard, Tracy Crane

Data Manager: Hagan Franks, Sarah Jane Wright, Damian Yukio Romero Diaz

Project Administrator: Steven Bethard, Tracy Crane, Sarah Jane Wright

Contributor: Rebecca Sharp, John Culnan, Damian Yukio Romero Diaz

Funder: National Institutes of Health (nih.gov)

Funding opportunity number: PAR-18-018

Grant: https://reporter.nih.gov/search/qfhaBJoM20qq64VSqwCScg/project-details/10109452

Template: DataWorks! Data Management and Sharing Plan Challenge

Project abstract:

Cancer survivors are a growing population in the United States; more than 16 million currently live in the US and by 2030 this number is expected to exceed 22 million. It is estimated that more than 50 percent of new cancer cases could be eliminated through a combination of healthy behaviors (e.g., physical activity and healthy diet); and cancer survivors are at high risk for developing new and

recurrent cancer. Unfortunately, a significant percentage of cancer survivors are not attaining the cancer preventive guidelines of healthy diet and physical activity. In the past few decades, a variety of telephone-based lifestyle interventions have demonstrated effectiveness in helping survivors meet cancer preventive guidelines, however these trials are labor intensive and expensive to deliver, limiting their potential for broad dissemination. We propose to address this hurdle by taking advantage of recent advances in artificial intelligence to reduce the cost and maximize the impact of these much-needed interventions. Machine learning (ML) and Natural Language Processing (NLP) are analytical techniques that automatically learn from direct and indirect patterns in data. We propose to use machine learned algorithms to analyze speech to aid in predicting who may be at risk of poor adoption of healthy lifestyle behaviors. These speech data will come from the Lifestyle Intervention for Ovarian cancer Enhanced Survival (LIVES) study, a telephone-based lifestyle intervention testing whether a diet low in fat and high in vegetables, fruit, and fiber, coupled with increased physical activity will increase time to disease progression in 1200 ovarian cancer survivors who have recently completed treatment, as compared to an attention control. Intervention coaches employed motivational interviewing to elicit behavior change and all calls on the LIVES trial were recorded with repeat assessments of diet, physical activity, patient reported and clinical outcomes. We will use this existing and robust longitudinal data set, which pairs conversational speech data with explicit outcomes, to achieve the following objectives. 1) Develop a ML model to identify patterns in the interactions between coaches and their participants that signal a likelihood of optimal behavior change in diet and physical activity given the comprehensive LIVES data set, utilizing voice recorded calls, demographics, and clinical and patient reported outcomes collected at multiple time points. 2) Decompose the ML model in terms of "intervenable factors", so that participant affect, coach adherence to the intervention protocol, and other important aspects of the interaction can be individually evaluated for their role in predicting behavior change, as well as adherence to intervention goals. This decomposition will directly enable early and targeted adjustments to intervention plans for individuals, reducing the cost and increasing the efficacy of intervention strategies. ML and NLP methods can produce models that listen to a coaching conversation and automatically predict whether it will result in positive change towards enactment of healthy lifestyle behaviors. Such predictive models would enable more efficient, effective, and individualized lifestyle interventions, the first step towards personalized behavioral medicine.

Start date: 12-31-2020

End date: 12-30-2022

Last modified: 02-20-2022

Copyright information:

Using natural language processing to determine predictors of healthy diet and physical activity behavior change in ovarian cancer survivors

Data Type

A general summary of the types and estimated amount of scientific data to be generated and/or used in the research. Describe data in general terms that address the type and amount/size of scientific data expected to be collected and used in the project (e.g., 256-channel EEG data and fMRI images from ~50 research participants). Descriptions may indicate the data modality (e.g., imaging, genomic, mobile, survey), level of aggregation (e.g., individual, aggregated, summarized), and/or the degree of data processing that has occurred (i.e., how raw or processed the data will be)

Used data

We use human subjects data including 1,048 telephone audio recordings (WAV), associated metadata (JSON) and patient demographics, clinical reported outcomes, and patient-reported outcomes (Microsoft EXCEL) from ~173 research participants engaging in Motivational Interviewing (MI) coaching from the previous Lifestyle Intervention for oVarian cancer Enhanced Survival (LIvES), GOG 0225.

Generated data (research outputs)

Ref.*	Description	Number**	Scale**	Data type	Preserved	Shared
A)	300-second snippets obtained from the LIvES telephone calls. Used for personality analysis	400	1GB	MP3	5 years	No
B)	Numerical responses from ~4 annotator's personality perceptions obtained from (A)	500	200MB	CSV	5 years	No
C)	Speaker-turn annotations from ~6 annotators of a subset of the LIvES telephone calls	90	200MB	JSON	5 years	No
D)	Aggregated speaker-turn annotations for inter-annotator agreement analysis	50	50MB	RTTM	5 years	No
E)	Linguistic and call content annotations of a subset of the LIvES telephone calls	85	450MB	JSON	5 years	No
F)	Transcriptions of the LIvES telephone calls	1,048	400MB	text	5 years	No
G)	Machine-learned models for processing linguistic data (automatic annotation)	6	18GB	binary or .tflite	10 years	Yes
H)	machine-learned models for processing predicting patient outcomes	4	12GB	binary or .tflite	10 years	Yes
I)	Computer code for the creation of machine-learned models (G) and (H)	10	200MB	Python files	10 years	Yes

* Reference (this will be used throughout this document), ** Expected

A description of which scientific data from the project will be preserved and shared.

The original data described above under "used data" is stored, shared, and managed by the LIvES study. We have obtained this data under the following terms: keep data secure, do not share data publicly. Our project will cease to use them after the project is finished in December 2022. The outputs (A-F) will be preserved for internal use for 5 more years (December 2022- December 2027), but will not be shared because we do not have participant consent for doing so. The resulting machine-learned models (G-H) will be preserved and shared openly in a data repository for at least 10 years. These models will not contain personally identifiable information or any health or health-related records.

A brief listing of 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.

We will make accessible all the computer code used to generate the machine-learned models (I) and the necessary documentation to use the computer code and models. No other metadata will be made accessible.

Related Tools, Software and/or Code

An indication of whether specialized tools are needed to access or manipulate shared scientific data to support replication or reuse, and name(s) of the needed tool(s) and software.

The machine-learned models will be distributed for their use with the Python (v.3.7+) programming language and will require one of the following free Python machine learning libraries (final decision pending): PyTorch (v.1.10.2) or TensorFlow (v.2.8.0).

If applicable, specify how needed tools can be accessed, (e.g., open-source and freely available, generally available for a fee in the marketplace, available only from the research team) and, if known, whether such tools are likely to remain available for as long as the scientific data remain available.

Python (https://www.python.org/), PyTorch (https://pytorch.org/), and TensorFlow (https://www.tensorflow.org/) are all freely accessible for most modern computers and at the moment there are no plans to discontinue their support. More information is available at their respective websites.

Standards

An indication of what standards will be applied to the scientific data and associated metadata (i.e., data formats, data dictionaries, data identifiers, definitions, unique identifiers, and other data documentation).

Data and metadata formats

- Microsoft EXCEL format is used for patient demographics, clinical reported outcomes, and patient-reported outcomes. We prefer this because these files are accessed often for consultation and EXCEL offers a more flexible interface than other table formats such as CSV.
- WAV format is used for storing LIvES telephone audio files.
- MP3 format is used for storing audio snippets (A). We use this format to save space and accessing time.
- JSON format is used for phone interview metadata, speaker-turn annotations, and linguistic and call content annotations (C) and (E).
- CSV format is used to store the ~500 numerical responses of the personality analysis and their corresponding metadata (B).
- The inter-annotator agreement analysis of speaker turns (D) usually makes use of the Rich

Transcription Time Marked (RTTM) metadata format. These files are space-delimited text files containing fields that are described in "The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan" (Appendix A) and include speaker, recording time, and recording information. The document can be found at: https://web.archive.org/web/20170119114252/http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf

- Python format (UTF8-encoded text files with .py extensions) is used for computer code (I)
- The machine-learned models will be stored and distributed in either binary (PyTorch) or tflite (TensorFlow) format, their associated processing files (vocabulary, tokenizer, etc.) will be stored in JSON format, and their corresponding documentation will be stored in Markdown format.

Other data

Data dictionaries for the original LIvES data are stored online in the UA REDCap instance and can be exported as CSV files if necessary. Any data dictionaries resulting from the current study are stored as CSV files in UA Box Health. Data identifiers and unique identifiers such as participant IDs are stored in encrypted Microsoft EXCEL files in the HIPAA-compliant cloud storage UA Box Health. Variable names and definitions for patient demographics, clinical reported outcomes, and patient-reported outcomes are stored in a Microsoft EXCEL file in UA Box Health.

Data Preservation, Access, and Associated Timelines

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

The computer code, machine-learned models, and documentation will be made publicly available through the University of Arizona Research Data Repository (ReDATA) and through the HuggingFace website (widely used within the machine-learning community) under an Apache 2.0 License. Although there are NRG-oncology approved repositories, most of them deal with specimens and images and none with machine-learned models of language interactions and so none of them were a good fit for this project.

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

The machine-learned models and associated metadata will be findable on ReDATA through a Digital Object Identifier (DOI) as well as a search engine at https://arizona.figshare.com/. It will also be findable through the search engine at https://huggingface.co/models

When the scientific data will be made available to other users (i.e., the larger research community, institutions, and/or the broader public) and for how long.

The machine-learned models and code will be available from December 2022 and for at least 10 years (as guaranteed by ReDATA). We cannot guarantee availability on the other two repositories.

Access, Distribution, or Reuse Considerations

Describe any applicable factors affecting subsequent access, distribution, or reuse of scientific data related to whether access to scientific data derived from humans will be controlled (i.e., made available by a data repository only after approval).

In accordance with the terms of our IRB-approved research plan, we will only share the resulting ~10 machine-learned models and associated computer code (G-I). We'll ensure that we create models where no HIPAA or other text information is recoverable. The models will be made publicly available under the Apache 2.0 license with no restrictions on academic or commercial re-use.

Oversight of Data Management and Sharing

Indicate how compliance with the Plan will be monitored and managed, frequency of oversight, and by whom (e.g., titles, roles).

The PIs for this project, Tracy Crane and Steven Bethard, will ensure that the data management plan is followed by auditing the project personnel on a monthly basis and monitoring the project through an online project management tool (Trello). Sarah Jane Wright is one of the data liaisons between the LIvES project and the current project. She is in charge of the patient outcome data, REDCap, questionnaire data, and patient personal records and identifiers. Sarah ensures that sensitive data is accessed on a case-by-case basis in a secure way through REDCap. Hagan Franks is the second data liaison between the LIvES project and the current project. He is in charge of the original audio telephone recordings. Hagan ensures that sensitive data is accessed on a case-by-case basis in a secure way through UA Box Health. John Culnan and Damian Romero are in charge of data annotation management and annotated files storage and safe-keeping. They provide University of Arizona approved annotators (number = 6) with HIPAA-compliant data for annotation, which is required for training supervised machine learning models. They are also in charge of creating machine-learned models. Steven Bethard is responsible for the overview of the machine-learned models. He is in charge of ensuring that the machine-learned models are sufficiently useful for future researchers and HIPAA compliant.

Planned Research Outputs

Software - "Python code for the creation of machine-learned models"

We will make accessible all the computer code used to generate the machine-learned models and the necessary documentation to use the computer code and models. This code can be used by researchers who have access to similar data (annotated patient telephone coaching recordings) to create machine-learned models that can automatically annotate and predict patient outcomes of the same kind. The computer code will be distributed as standard Python (.py) UTF8-encoded files.

Software - "Machine-learned models for data annotation"

We expect to release ~6 machine-learned models for data annotation in one of two open formats (to be determined): PyTorch or TensorFlow. These models will help researchers automatically annotate similar data with some accuracy. While the results may not be completely accurate, they will be provided for future researchers to help their language data annotation efforts, which may reduce the economic impact of their project. We expect the models to be able to do the following functions: divide speaker turns according to raw audio from telephone motivational interviews; divide speaker turns based on text transcriptions and time-stamps derived from motivational interviews, detect linguistic constructs such as questions based on text transcriptions; and profile speaker's personality based on raw audio from motivational interviews.

Software - "Machine-learned models for outcome prediction"

We expect to release ~4 machine-learned models for outcome prediction in one of two open formats (to be determined): PyTorch or TensorFlow. These models will help researchers automatically predict dietary outcomes for patients undergoing motivational interview interventions. There will be at least one general model that will identify patterns in the interactions between coaches and their participants that signal a likelihood of optimal behavior change in diet and physical activity given voice-recorded motivational interviews. There will be at least two different models of different "intervenable factors" that will identify participant affect, coach adherence to the intervention protocol, and other important aspects of the interaction that will be individually evaluated for their role in predicting behavior change, as well as adherence to intervention goals.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Python code for the creation of machine-learned mo ...	Software	2022-12-30	Open	ReDATA		Apache License 2.0	None specified	No	No
Machine-learned models for data annotation	Software	2022-12-30	Open	ReDATA		Apache License 2.0	None specified	No	No
Machine-learned models for outcome prediction	Software	2022-12-30	Open	ReDATA		Apache License 2.0	None specified	No	No