Digital Phenotyping and Cardiovascular Health

NCT04574882 (IRB #: 833699)

December 7, 2021

Final Analysis Plan

As digital data exist in multiple formats (e.g. posts, photos, searches, steps) we will analyze each source according to its specific attributes. We will systematically apply robust data-driven approaches to parsing and stratifying data. This step requires significant data pre-processing and annotation as raw digital data is not directly interpretable. This will also occur over multiple iterations over the course of the grant as new terms (e.g., the term "vaping" describes use of e-cigarettes) and digital sources emerge. To automatically extract meaningful variables from the text, we will use both dictionary-based (e.g., counting words from previously constructed word lists, lexica) and data-driven "open vocabulary" methods which automatically generate lexica and identify important words and phrases based on the data itself rather than a priori categories. For dictionary analyses, we will use the Consumer Health Vocabulary, an online open source thesaurus of colloquial and technical medical terms and the Unified Medical Language System which contains lists of medical terms and vocabulary to identify disease discussions. We will specifically focus on words and words associated with previously identified predictors of CHD (e.g. demographics, family history, lifestyle, comorbidities, blood pressure) as identified in a systematic review of prediction models for cardiovascular risk by Damen et al. BMJ 2016. We will create classifiers to determine if a given instance of a disease term is referring to the disease or something unrelated (e.g. "If I see a celebrity I will have a heart attack" vs. "I am having chest pain, SOB and had a prior heart attack"). Our prior work has demonstrated that, as one might guess, some CV diseases, like "hypertension", are almost always used in their medical context while others, like "heart attack", are often used colloquially. We will also create classifiers to categorize language by broad grouping such as risk factors, treatment, medication adherence, symptoms, and by more granular categories such as active/sedentary, healthy food/unhealthy food, smoking. We will compare the performance of several machine-learning algorithms (e.g. support vector machines, random forests, and, in cases with adequate data, deep learning) and assess the accuracy of our disease term classifier against hand-annotated data. For multiword phrases we will use approaches for determining the probability that words occur together (e.g., pointwise mutual information PMI). We will also use additional linguistic models to account for more subtle language complexities and nuance such as parts of speech, and disambiguated words. We will use automated computational methods (e.g., deep learning-based object recognition) combined with manual coding to classify features (e.g., healthy or unhealthy food, physical activity, smoking, and drinking) in images. To determine nutritional content of food images, for example, we will use the imagga.com image classification software, combined with analysis of the words used in the text associated with images to recognize food images. Those images that are food images will then, to the best of our ability, be compared to a list of healthy and unhealthy

groupings (e.g. high salt, high fat) and also compared to a USDA database of precise nutritional values for over 30 nutrients for 8600 food items. We will also construct a convolutional neural network based on labeled food and health behavior-related images to train a model to recognize images in the most frequent categories which our first pass fails to find. We will analyze the labeled images to better understand the nature of food posting, and how they vary with the time of day, demographics of the person posting the image We will compare summary data from smartphones with self-reported data. Time series data from sensors will be used to analyze presumed changes in behavior using a previously validated health behavior activity change detection framework which segments time series data by time periods, identifies changes and their significance, and identifies the presumed locus of changes. The strength of using different types of data (text, images, sensors) is that each provides a unique type of information. Sensors provide direct measurements of protective and risk factors for CV, but they only measure what they are intended to measure. On the other hand, text and image data have been shown to capture a wide variety of information about people from demographics and personality to mood and beliefs. Further, each patient may use differing proportions of each modality such that, for example, text may make up for a lack of images and vice-versa. As digital media data may over-represent, under- represent or not represent health behaviors we will compare survey data and clinical data with digital data to quantify differences in behaviors and risk factors. We will explore statistical and machine learning methods for combining heterogeneous variables into predictive models. For our main analysis, we will both explore the correlations between the different modalities to better understand when they provide redundant or non-redundant information and we will, more importantly, combine the different modalities into a single predictive model. Standard machine learning methods such as random forests make it easy to add heterogeneous features to a single predictive model, and give measures of variable significance' indicating how much each feature contributes to predictive accuracy. We will also test more sophisticated machine learning methods for combining multiple modalities in a single model, such as using extensions of the 'elastic net' regression and dimensionality reduction framework that account for the fact that different modalities require different regularization penalties. We will also explore imputation and "back-off" methods that account for the fact that different patients will have substantially different amounts of data available in different modalities. These methods work by either estimating ("imputing") values for the missing data or automatically adjusting the weights given to different modalities for each patient based on the amount of data available for that patient. For Objective 1: Sample characterization: We will use summary statistics to describe patient demographics, survey response data, and digital data usage (e.g. frequency of posting, number of social media platforms). Descriptive data will be presented as mean (SD) for continuous variables and frequencies of participants for categorical variables. 5b Extracting topics and features: We will use the Mallet package for implementation of the Latent Dirichlet Allocation (LDA) method for clustering language data of all participants. The LDA probabilistic model assumes that documents (e.g. Facebook status updates) contain a distribution of topics which then contain a distribution of words. Ultimately, words are grouped together by considering the other words they appear with. Descriptive statistics, t-tests (continuous measures) and chi-squared tests (categorical measures) will be used to compare topics and features with survey data. For example, NHANES survey questions (e.g. Do you now smoke cigarettes?) would be compared with images of cigarette smoking. For Objective 2: We will compare prediction models with and without digital variables. Probability weighted Cox proportional hazard analysis with robust variance estimates will be used to assess the association between each variable from digital media and CHD in univariable and multivariable models. We will use graphical and analytical methods to test the proportional hazards assumption for each analysis. To evaluate the predictive strength of digital data, we will build four predictive machine learning models using a stepwise sequence to select the most parsimonious model using combinations of risk factors. (1. Framingham predictors / American College of Cardiology ASCVD Risk Estimator Plus alone, 2. Framingham predictors / American College of Cardiology ASCVD Risk Estimator Plus + SES, 3. Framingham predictors / American College of Cardiology ASCVD Risk Estimator Plus + digital topics/ features, 4. Framingham predictors / American College of Cardiology ASCVD Risk Estimator Plus + digital topics/features+ SES) For Objective 3: We will use Super Learner, the ensemble machine learning approach which allows for the specification of multiple plausible candidate prediction models. Each of the candidate models is applied to a training set and outcomes are predicted using the validation set. A loss function is calculated within each validation set and then averaged across validation sets, which provides the estimated cross-validated risk score for each method. Calibration will be assessed using a calibration plot comparing the predicted probabilities (obtained using fixed regression coefficients of the prediction models) with the observed probabilities. The Super Learner algorithm finds the optimal weighted combination across all of the specified methods. Van der Laan et al. proved the asymptotic efficiency of the Super Learner algorithm and demonstrated that the optimal combination performs at least as well as the best estimators from the candidate models. We will implement Super Learner in R (The R Foundation for Statistical Computing). In this approach generalized estimating equation (GEE) and generalized linear mixed effects models will be explored to account for repeated measures per patient over time. One of the challenges of cost analyses is censoring, since cost may be incomplete and available for some patients and not others. As the dependent variable will be cost, we will be attentive to informative censoring and right skewness, potentially using generalized linear models under a gamma distribution and log link. To account for informative sensoring we will use inverse probability of sensoring weighting. We will also explore the application of our new nested g-computation approach that allows for informative censoring and time- varying covariates to obtain future cost predictions. Additionally, we will generate summary statistics to describe the cohort and survey responses. We will use the model-based approach of latent class analysis (LCA)- to classify individuals into previously unmeasured subgroups. Variables to include in the LCA will include: demographics, years of platform use, posting: volume, quality, retransmission, and engagement. We will use the Bayesian information criterion to assess goodness of model fit, entropy to assess variations between classes, and the parametric bootstrapped likelihood ratio test to assess if a model with k classes has better performance than k-1 classes. We will then use a regression model to compare with measures of health status, social media perceptions, and readiness/activation for each latent class. Models will be adjusted for demographics.