REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
01-02-2019	Final Report	1-Nov-2013 - 30-Apr-2014

4. TITLE AND SUBTITLE

Final Report: Building a next generation Model for Biomedical Research: Validation of Health Sensors using Online Community Registries and Collaborative Data Interpretation

5a. CONTRACT NUMBER

W911NF-14-1-0002

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES

Sage Bionetworks
1100 Fairview Ave North
MS: M1-C108
Seattle, WA 98109 -1024

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)

ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

64867-LS-DRP.1

12. DISTRIBUTION AVAILIBILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Stephen Friend
UU	UU	UU	UU		19b. TELEPHONE NUMBER: 206-667-2101

RPPR Final Report
as of 28-Feb-2019

Agency Code:

Proposal Number: 64867LSDRP **Agreement Number: W911NF-14-1-0002**
INVESTIGATOR(S):

Name: Stephen H Friend
Email: friend@sagebase.org
Phone Number: 2066672101
Principal: Y

Organization: **Sage Bionetworks**
Address: 1100 Fairview Ave North, Seattle, WA 981091024
Country: USA
DUNS Number: 830977117 EIN: 264489946

Report Date: 31-Jul-2017 Date Received: 01-Feb-2019
Final Report for Period Beginning 01-Nov-2013 and Ending 30-Apr-2014
Title: Building a next generation Model for Biomedical Research: Validation of Health Sensors using Online Community Registries and Collaborative Data Interpretation
Begin Performance Period: 01-Nov-2013 **End Performance Period:** 30-Apr-2014
Report Term: 0-Other

Submitted By: Amy Truong Email: amy.truong@sagebionetworks.org
Phone: (000) 000-0000

Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: **STEM Participants:**

Major Goals: Objective(s):
- Identify issues and challenges with preparing open data collaborative crowd interpretation such as the privacy and ethical implications (e.g. re-identification) [Tasks 1-2]
- Assess feasibility of gathering voice samples from Parkinson's disease patients over the Internet as a marker of disease severity including data quality, predictive power, error using a seedling pilot sample of N=500 [Tasks 3]
- Validate ability to predict specific elements of impact / disability of disease e.g. tremor, slowness, activities of daily living in seedling pilot [Task 4]
- Pilot a collaborative crowd-sourced data competition as a "dry run" [Task 5-7-8]
- Report characteristics of engaged users relative to community / trial PD population [Task 6]

The study is designed to (1) rapidly collect under appropriate permissions voice data recordings "in the wild" through regular land or cell phone lines and responses to a standard health questionnaire from 500 volunteers with Parkinson's Disease; and (2) to invite data analysis experts to collaboratively to predict the self-reported disease severity from the voice features.

Specific Aims
1) Collect recordings of sustained voice phonations matched with Parkinson's disease self-assessment questionnaire (PDRS) and limited de-identified phenotypic covariates from PatientsLikeMe Parkinson's Disease community volunteers.
2) Assess the quality and predictive power of voice recordings "in the wild" for optimization.
3) Invite selected analysts to collaboratively determine the best methods to correlate the severity of self-reported symptoms to the voice pattern characteristics.
4) Establish the conditions to scaling up this effort, creating a dynamic patient sensor network and deploying a wider open analysis challenge in a future project phase.

Accomplishments: See attached PDF titled Final Report-Patient Voice Analysis

Training Opportunities: Nothing to Report

RPPR Final Report
as of 28-Feb-2019

Results Dissemination: Nothing to Report

Honors and Awards: Nothing to Report

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: Co-Investigator
Participant: Arno Klein
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Other (specify)
Participant: Christine Fabre Suver
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Project Report

Building a next-generation model for biomedical research: validation of health sensors using online community registries and collaborative data interpretation

Contributors by first name alphabetical order: Arno Klein*, Ben Heywood#, Christine Suver*, Elias Chaibub Neto*, Marcy Fitz-Randolph#, Max Little#, Paul Wicks#, and Stephen Friend*.

* Sage Bionetworks, # PatientsLikeMe

Short title: Patient Voice Analysis
Proposal Log Number 64867-LS-DRP,
Award Number W911NF-14-1-0002,
HRPO Log Number A-18007

Executive Summary

The current siloed approach to biomedical research is ill equipped to take advantage of the emerging tools to query biological systems and remarkable data now available. Our original proposal was for a pilot study to assess the feasibility of using a community-based collaborative approach to doing research, engaging individuals with Parkinson's disease, and monitoring diseases using voice pattern recognition. This project establishes a scientific challenge to optimize the analysis of self-reported outcome data and sustained voice phonation recordings of individuals from the PatientsLikeMe Parkinson's disease online patient community. The main goal of the project is to establish through collaborative data interpretation whether non-lab quality voice recordings such as web-based, cell-phone or home phone recordings could function as valid health sensors that can be used to predict patient function and disease severity. The relationship between aspects of voice such as pitch, volume and timbre, and characteristics of Parkinson's disease such as tremor or slowness, might be complex and multifactorial. Therefore we invited cross-disciplinary experts to develop algorithms to determine disease severity from these data. The underlying goal was to assess the therapeutic value of voice analysis. Another goal for this project was to identify the barriers and solutions to expand the scope of this work to broader community engagement, longitudinal measurements, smartphone integration and modeling of therapeutic drug effects. This was a feasibility study, intended to precede the design of a larger-scale project.

May 2014

Table of Content

Acronyms

H&Y: Hoehn & Yahr staging
PD: Parkinson's disease
PDRS: Parkinson's Disease Rating Scale
PLM: PatientsLikeMe
PVA: Parkinson's Voice Analysis
PVI: Patients Voice Initiative
SB: Sage Bionetworks
UPDRS: Unified Parkinson's Disease Rating Scale
WIRB: Western Institutional Review Board

Background and Objectives

Parkinson's disease (PD) is a relatively common, progressive neurological disorder affecting approximately 0.3% of the general population in industrialized countries. It generally affects people over 60 years, but rarely can strike younger people diagnosed under the age of 40. PD is considered a movement disorder (i.e., it affects the ability to perform normal voluntary motion), but many patients also experience cognitive impairment and emotional/mood disturbances. The classic movement symptoms of PD include a 4hz tremor, rigidity throughout the body, and slow or hesitant motion. These movement problems often have a substantial negative impact on the ability of the patient to perform essential everyday activities of daily living, such as bathing, dressing, turning in bed, walking unaided, and getting up from a sitting position. The cause of PD is currently thought to be the loss of dopaminergic neurons in an area of the brain known as the substantia nigra. PD is incurable, and there are no conclusive diagnostic tests; it is a clinical diagnosis of exclusion. The most accurate diagnosis based on behavioral symptoms achieves, at best, 90% accuracy when compared to postmortem pathological examination.

Management of PD is complex and relies on an expert multidisciplinary team. In the ideal situation patients are regularly seeing a neurologist trained in movement disorders who can assess their function, infer their likely degree of dopamine depletion, and replace this with dopamine precursors, agonists, and metabolism inhibitors while balancing out the side effects these drugs can have and making referrals to physiotherapists, speech and language therapists, and occupational therapists. In a well-run movement disorder clinic, patients would also have access to nurse-lead symptom management for pain, insomnia, and depression. Some more advanced clinics, or those who conduct a lot of clinical research may use validated tools such as the Hoehn & Yahr (H&Y) Scale or the UPDRS. The H&Y is a coarse quantitative ordinal measure of symptoms that assigns a number from 0 to 5, with 0 being healthy and 5 denoting severe disability[1]. Despite the benefits of frequent measurement and multi-disciplinary care, more than 40% of people suffering from Parkinson's disease have never even seen a neurologist, let alone a movement disorders specialist [2]. Where more sensitive measurement is needed, such as in clinical trials, H&Y has been largely supplanted by the ordinal Unified Parkinson's Disease Rating Scale (UPDRS) (version 3.0)[3] or the Movement Disorders Society UPDRS (MDS-UPDRS)[4], which are more time consuming (and therefore expensive) to administer but are more precise.

Table 1: Hoehn & Yahr (H&Y) Staging system

Stage	Description
0	No signs of disease
1	Symptoms on one side only (unilateral)
1.5	Symptoms unilateral and also involving the neck and spine
2	Symptoms on both sides (bilateral) but no impairment of balance
2.5	Mild bilateral symptoms with recovery when the 'pull' test is given (the doctor stands behind the person and asks them to maintain their balance when pulled backwards)
3	Balance impairment. Mild to moderate disease. Physically independent
4	Severe disability, but still able to walk or stand unassisted
5	Needing a wheelchair or bedridden unless assisted

UPDRS values have been collected on individuals at all stages of the disease in clinical trials, and there is substantial research data available on PD symptom progression quantified on this scale (e.g. DATATOP [5]). This kind of data has been used to calibrate models of PD symptom progression over the course of years to decades. However, the full UPDRS is a complex test that requires expertise to administer, attendance of the patient in the clinic, and the average time for administration of the full test is approximately 17 minutes [6]. Unfortunately, these difficulties mean that it is usually prohibitive to objectively score PD severity on timescales shorter than 3 months (low frequency). Since most longitudinal UPDRS data is low frequency, objective information about symptom dynamics occurring on a shorter timescale than 3 months (high frequency data) is lacking.

Advocates of telemedicine have proposed that such shortcomings could be addressed through the use of web-based technologies such as webcams to enable a wider swathe of patients to access a movement disorder specialist [7]. In such settings, the neurological examination of H&Y and UPDRS would be achieved by the neurologist asking the patient a series of interview questions around activities of daily living and then asking them to perform specific tasks in front of the camera such as tapping their fingers together as quickly as possible; PD patients show a characteristic slowing of tapping ability particularly on the side contra-lateral to the site of worst dopamine loss. Initial studies show that such clinical examination is technically feasible remotely, has a high degree of patient acceptance, and could result in major efficiency savings while expanding access. However, these pilots have been limited in scope and the major roadblock appears to be the lack of reimbursement potential for telemedicine visits [8].

An alternative approach to this reliance on clinicians is instead to empower people with patient reported outcome (PRO) versions of validated rating scales like the UPDRS and H&Y. This is the approach taken by the patient powered research network (PPRN) PatientsLikeMe (PLM), which has recruited over 6,000 members with Parkinson's disease since 2007 and has produced over 30 peer-reviewed journal publications including an Internet-based clinical trial of lithium carbonate for ALS, the development of new measurement scales in neurological diseases, and a number of studies exploring Parkinson's disease [8–10]. The PD community on PLM permits members to submit their symptoms, treatments, and a self-reported patient reported outcome similar to that used in clinical trials: the Unified Parkinson's Disease Rating Scale (UPDRS). However a limitation of self-report is the inherent bias affecting recall,

assessment, and accuracy and objective measurement to supplement patient report would be preferable.

One novel technique to address these shortcomings is that proposed by the Parkinson's Voice Initiative (PVI), which has collected over 12,000 audio samples over a standard telephone line, about a third of whom have PD and the rest of whom do not. Through earlier work conducted in controlled lab settings, Dr. Max Little established that characteristic vocal patterns elicited in the PVI test can diagnose PD with a high degree of specificity and sensitivity. These initial results suggest that lab-quality digital audio recordings of voice phonation can be a surrogate marker of disease detection. The potential of cheap, fast, accurate and objective measurement of PD using any telephone seems a promising method to bridge the gulf between the dearth of measurement in the broader PD community and where the field needs to be in order to provide best quality care to patients, including those who lack access to a neurologist specializing in PD.

A logical extension to this work would be to establish whether voice characteristics might be correlated with the severity of an individual patient's PD over time. If validated, such a tool might allow the PD field to increase the frequency with which patients are monitored without the need for additional neurologists or unnecessary waste. It might be possible to better prioritize those individuals who need to see the care team more or less frequently, and to measure the impact of medication changes prescribed on an objective measure of function. However, to conduct such work would require that a large number of people with PD complete a validated disease measurement and provide a voice sample, and to conduct statistical analysis to identify any potentially relevant voice features.

We believe that the best approach towards developing robust and accurate predictive models of PD severity is to enable an open diverse community where data access is simple and people are incentivized to share their analysis methods and results. Prize-based data analysis challenges and/or competitions are very effective at soliciting contributions from skilled analysts. This is especially true for data analysis problems that can be solved using sophisticated machine learning and data mining methods. The main advantage of such open challenges lies in encouraging a diversity of analytical approaches from across scientific disciplines to solve inherently difficult but important questions. This approach has been suggested as a way to find new drugs[11] by soliciting contributions from a large online community. Indeed, Sage Bionetworks successfully used such open challenge approach in genomic analyses to predict breast cancer survival from clinical and omics data and is using this approach to identify genetic predictors of response to immunosuppressive therapy in individuals with Rheumatoid Arthritis [12–14].

In this study we combined Dr. Little's voice recording technology with the online platform from PLM to conduct rapid data collection, provided the anonymized data to a crowd-sourced data analysis platform, Synapse (Sage Bionetworks [15]) and initiated a distributed analysis with selected experts in a dry-run analysis Challenge to help assess the therapeutic value of voice pattern recognition.

Objective(s):

- Identify issues and challenges with preparing open data collaborative crowd interpretation such as the privacy and ethical implications (e.g. re-identification)[Tasks 1-2]

- Assess feasibility of gathering voice samples from Parkinson's disease patients over the Internet as a marker of disease severity including data quality, predictive power, error using a seedling pilot sample of N=500 [Tasks 3]
- Validate ability to predict specific elements of impact / disability of disease e.g. tremor, slowness, activities of daily living in seedling pilot [Task 4]
- Pilot a collaborative crowd-sourced data competition as a "dry run" [Task 5-7-8]
- Report characteristics of engaged users relative to community / trial PD population [Task 6]

Leadership Team

- Sage Bionetworks (Open consent, Synapse platform, distributed analytics)
    - Stephen Friend MD PhD. - President, co-founder and director of Sage Bionetworks and Ashoka Fellow. Strong advocate of open science. Formerly Senior Vice-President at Merck & Co. and co-founder of Rosetta Inpharmatics with Leland H. Hartwell and Leroy Hood.
- PatientsLikeMe (Online data capture, Parkinson's community & clinical expertise)
    - Paul Wicks, PhD. – Neuropsychologist with specialist training in Parkinson's disease, 30+ publications in online medical research, TED Fellow
    - Max Little, PhD. – Applied mathematician specializing in voice analysis in Parkinson's disease, 30+ publications in voice analysis, TED Fellow
    - Ben Heywood – Co-founded PatientsLikeMe. Advocate of "openness policy" concerning health data for analytics[16].

Study Design and specific aims

The study is designed to (1) rapidly collect under appropriate permissions voice data recordings "in the wild" through regular land or cell phone lines and responses to a standard health questionnaire from 500 volunteers with Parkinson's Disease; and (2) to invite data analysis experts to collaboratively to predict the self-reported disease severity from the voice features.

Specific Aims

1) Collect recordings of sustained voice phonations matched with Parkinson's disease self-assessment questionnaire (PDRS) and limited de-identified phenotypic covariates from PatientsLikeMe Parkinson's Disease community volunteers.
2) Assess the quality and predictive power of voice recordings "in the wild" for optimization.
3) Invite selected analysts to collaboratively determine the best methods to correlate the severity of self-reported symptoms to the voice pattern characteristics.
4) Establish the conditions to scaling up this effort, creating a dynamic patient sensor network and deploying a wider open analysis challenge in a future project phase.

Ethical & Regulatory Reviews

This project has two arms: (a) Data collection through PatientsLikeMe (PLM), and (b) Analysis challenge organized by Sage Bionetworks (Sage).

PLM and Sage submitted their respective arm of the project to Western Institutional Review Board (WIRB) as well as to the US Army Medical Research and Materiel Command (USAMRMC), Office of Research Protections (ORP), Human Research Protection Office (HRPO) for independent ethical and regulatory review. Both arms of the project were found to comply with applicable WIRB, DOD, US Army, and USAMRMC human subjects protection requirements. It was designated as not posing greater than minimal risk.

PLM uses an online process to recruit participants, provide information about the project and obtain consent. WIRB found that information presented by PLM on its site met the elements of informed consent and granted a "Waiver of documentation of consent" since consent would not be in the form of a traditional signed consent form.

The voice data collected in the study does not represent identifiable speech recordings.
Biometric data, such as voice recordings, could be identifiable and are considered Protected Health Information according to HIPAA guidelines. In this project however, because the recordings are brief (less than 30 seconds) and consist only of sustained phonations (individual making the vowel sound 'aaah'), it is unlikely that they could be used to unambiguously re-identify participants. The current scientific consensus in speaker identification, which is the discipline concerned with the problem of identifying individuals from speech recordings, is that this voice data would fall far below the minimum data requirement in order to identify individuals [17] . There are several reasons for this, but the most important are:

(a) Speech data is required, that is, we must have examples of identifiable words. By contrast, the data in this study is of the single sustained phonation 'aaah' alone,
(b) Several minutes of example speech from each individual is required. In this study, we only have at most 30 seconds' of voice data from each individual.

Once both PDRS and voice data were collected, PLM ensured that information that could directly identify the participant was removed and replaced by unique random identifiers that cannot be used to easily ascertain participant's name. Sage Bionetworks received only the de-identified coded data for posting on Synapse.

Collecting the voice phonation and PDRS were not found to increase the risk for re-identification of the participants. Nevertheless, to remove any ambiguity WIRB board approved Sage Bionetworks' request for a WAIVER OF CONSENT AND WAIVER OF AUTHORIZATION for use and disclosure of protected health information (PHI) for this project, allowing sharing of the collected data through its analysis platform called Synapse (Synapse, https://www.synapse.org). Data distribution through Synapse is governed by a set of procedures and principles designed to meet legal, ethical and regulatory standards for the sharing of human data (WIRB Protocol 20112068). The IRB-approved governance process includes well documented Terms and Conditions of Use, guidelines and operating procedures, privacy enhancing technologies, quarterly audits of data use and annual reviews of the governance system by a panel of worldwide experts in bioethics.

The initial protocol focused on doing a modest analysis in a "dry-run" to establish baseline predictions. Yet, participants contributed their voices and associated data on the general understanding that their data will be used more broadly to promote progress in scientific research. WIRB found that there is no meaningful increased risk by sharing the de-identified coded data on Synapse beyond the dry-run analysis phase (WIRB-20112068- Amendment1).

Data Description

The data include brief recordings of sustained phonation, self-assessed disease scale, and self-reported outcomes from volunteers with Parkinson's disease. The participants in this study are members of the PatientsLikeMe (PLM) Parkinson's disease community who voluntarily provide the information to the PLM platform to broadly share for research (http://www.patientslikeme.com/).

PLM is an online data-driven research platform for people who wish to share their health experiences and connect with other participants PLM collects and graphs self-reported health information from participants and provides them with a context in which to manage and monitor their condition so they can learn to take control of their health. PLM has established a Parkinson's disease community of about 6,000 individuals with Parkinson's disease. PLM operates under an openness philosophy well documented in their User Agreement and Privacy Policy.

Parkinson's disease Rating Scale (PDRS)

PDRS is an abbreviated version of the Unified Parkinson's Disease Rating Scale (UPDRS) widely used by PLM to evaluate PD severity in clinical trials. The 18 items PDRS questionnaire omits the clinical observation section present in the more comprehensive 42 items UPDRS. PDRS can be self-administered and completed quickly (~ 10 minutes). PDRS provides a maximum score of 72 points. Each question is rated on a (0-4) scale with "0" representing no disability and "4" worst disability. No protected health information is included in the PDRS.

Phenotypic covariates

A limited set of covariates was also captured: participant's age and gender, years since first symptoms and whether the PDRS response reflected performance with or without PD treatment. Time and date of completion of the PDRS and voice recordings were automatically recorded as well.

Voice characteristics

Voice recordings were collected using proprietary code running on the Twilio IVR system. As of the initiation of the dry-run, over 800 recordings were collected. These consist of up to 30 seconds of the simple sustained phonation 'aaaah' provided by each participant. From analysis of the call logs, it is apparent that no calls were dropped and all were successfully captured by the IVR system.

Data Collection, Processing and QC

The project aimed to collect data from 500 individuals with PD over the course of eight weeks (January and February 2014). There were four main requirements of each patient for this project:

1. Complete PDRS and H&Y scale on the PLM site
2. Read the PVA project information page within PLM and agree to participate
3. Complete the voice recording
4. Enter the recording reference number -given at the end of the recording session- onto the PVA PLM page.

Completing the PDRS

Every member of the PLM Parkinson's Disease Community can complete the Parkinson's Disease Rating Scale (PDRS) survey and determine their H & Y scale score on the PLM site whenever they desire. They are also asked whether their response reflects how they feel with or without taking their PD medication. At the bottom of the PDRS survey, in addition to the usual buttons for submitting the survey or canceling the activity a third prominently placed button was added to continue to the PVA project, and a fourth button to obtain more information about the PVA project. When selecting the information button, research FAQ information appeared explaining the research question and the steps required, as well as all information that correspond to the elements of informed consent.

Reading the PVA project Information page

Clicking on the Continue button at the bottom of the PDRS page took PLM members to the main PVA project page. This page included the instructions for the patient regarding the voice data collection, an entry box for recording a reference number given at the end of the voice recording session, and a sidebar of information about the project, including information that form the elements of informed consent. The placement of this information ensured that all participants had the project research information visible while they were participating.

The instructions directed willing participants to the phone numbers to call in order to record their voice data. Toll-free telephone numbers were available for the United States and Canada; there were also numbers (not toll-free) for the UK, Australia and New Zealand. Participants were asked to have pen and paper ready to write down their reference number at the end of the call.

Completing the voice recordings and entering the reference number

Participants were asked to take a deep breath, and say 'aaah' for as long as they can, after the beep". An actual 'aaah' was demonstrated by the voice prompts as an example, and held at a steady pitch for a couple of seconds. At the end of the recording participants were given a seven-digit reference number that they had to enter on the PLM PVA page. If an incorrectly formatted number was entered, an error message would appear asking participants to check their entry

Quality Control

For this seedling project, only rudimentary quality control was applied to the data.

The time segments of the recording that contained sounds of sufficiently large amplitude for at least a few seconds were included for further analysis. No attempt to determine the nature and utility of these sounds were made. This quality control algorithm was derived from a similar algorithm demonstrated to be effective at processing high-quality lab recordings. However, since these are telephone quality recordings, the type I/II quality control error rates for this kind of algorithm are unknown.

The date and time stamps were used to select the PDRS submitted closest in time to the voice recording reference number. Because of the differences in the time zones of the voice server and the PLM site server, the data set construction criteria required choosing the closest PDRS

(at 0 or 1 days from voice). Also, incorrectly entered reference numbers were cross-referenced against the voice database to see if the correct number could be regenerated. If a patient entered the same reference number multiple times, the duplicates were removed from the data set.

Not all participants had complete covariates, but the PDRS and voice data were included to see what the effect of the missing data might be on the final analyses.

Finally, each patient was given a unique identifier. This allowed for identification of participants who provided more than one PDRS/voice data pair for the project; although each voice recording was given a different reference number, the unique identifier would identify them as the same person.

In total, over 850 reference numbers were entered by participants, in excess of the anticipated 500 originally scoped. After pairing with PDRS (within 1 day) and excluding duplicated and malformed or incorrect reference numbers, 818 PDRS records were uploaded to Sage.

De-identification

To preserve participant's privacy and data confidentiality, PLM removed all obvious information that could directly identify the participant prior to contributing data to Synapse. Therefore the data coded with unique random identifiers cannot be used to easily ascertain participant's name.

Voice Data Processing

For those segments of voice recordings that passed the quality control, a range of features (38 in total) were extracted. These ranged from pitch, amplitude and itch/amplitude stability-based to cepstral and spectral power measures. The Mel-Frequency Cepstrum coefficients are 'standard' MFCCs used in speech processing, generated as followed. First, the windowed power spectrum is computed for each frame (currently 20 msec). Then, the power spectrum is grouped together into logarithmically-spaced sets of frequency ranges (the mel scale). Next, the logarithm of these grouped log frequency bins is found, and the discrete cosine transform is applied to get the mel scale cepstra. Finally, the cepstra is filtered to get the final set of MFCCs.
This choice of features was shown to be effective for lab and home-based recordings collected using high quality recording equipment. As with quality control, it is unknown whether very poor telephone-based recordings would confound these features.

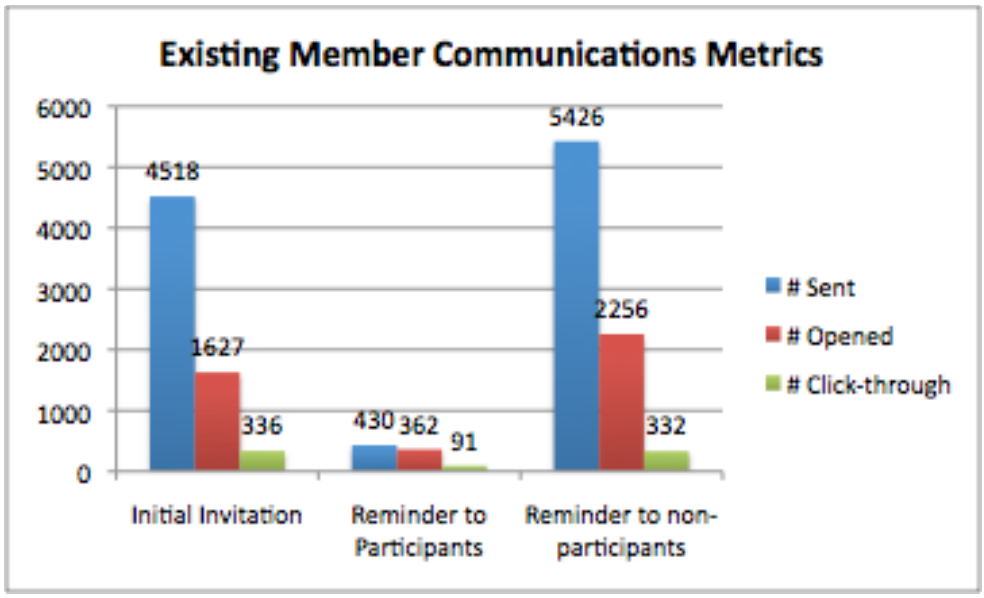
Outreach and recruitment

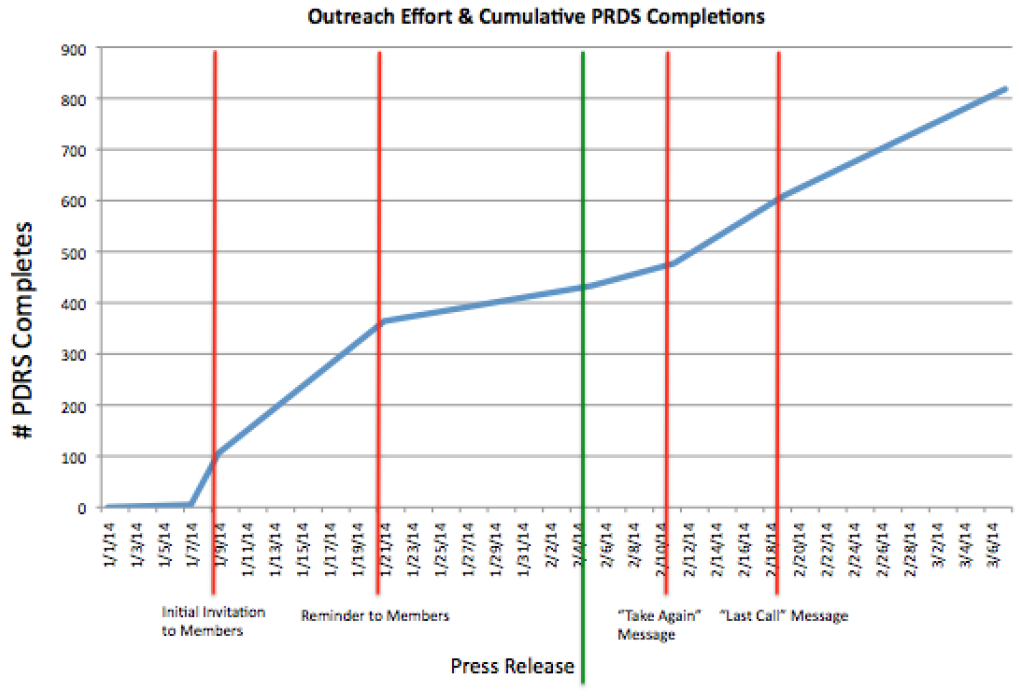
In order to maximize the number of participants successfully completing the rather complex data entry process required in this project, a series of communications were employed to reach out to both existing and new PLM members. The primary challenges in outreach and recruitment included raising awareness of the PVA project and guiding PLM members through the data collection processes. The user experience related to the data collection flow is detailed in the Data Collection section of this report.

Announcement of the project and recruitment began with the PLM PD community. Since PLM had previously worked with Dr. Max Little on his PVI project, many members of PLM PD community had some familiarity with the research questions behind this new endeavor. 4518 active PLM members with PD were sent messages announcing this new project with Dr. Little and asked to participate. The message included a link to the PDRS online survey, an existing feature of the PD community experience. The language in the invitation guided people through the steps of the project: complete the PDRS, call the voice recording number and complete the voice recording, and enter the reference number given at the end of the call (detailed in the Data Collection sections of this report).

Additional messaging to PLM members included a message in the middle of the eight week data collection period which reminded them that they could participate and submit PDRS-voice recording pairs more than once during the collection period; almost 100 participants contributed more than one time. Finally, another message was sent in the last week of data collection as an alert of the limited amount of time left to submit their voice recordings for the PVA project.

A blog post was also created for the PD community the steps to participate in the study as well as some background information from the previous PVI project with Dr. Max Little and more information about Sage Bionetworks.





In addition to outreach aimed at existing PLM members, communications were created to encourage new members to participate in the PVA project. The outreach campaign consisted of two major components: a joint press release with Sage Bionetworks, and a series of social media messages delivered via Facebook and Twitter. The outreach to new members focused on recruiting them to the new PVA "landing page," a page specifically set up for PD communities to join PLM for the purpose of participating in the project. The User Experience and Design he team at PLM created this landing page to inform and raise interest in the project. It featured a quote from Dr. Max Little and the steps required to participate in the study. The page also had a link to the same research information FAQ (elements of consent language) that appeared on the PVA project page.

The joint press release was created in cooperation with Sage Bionetworks and released on February 5, 2014. It was disseminated to organizations and picked up by numerous news media in the US and UK, as well as sparking discussion in the Parkinson's UK forums. Dr. Max Little also participated in outreach by discussing the project in a TEDMED blog. Individual PLM staff members also reposted the press release announcement on LinkedIn.

The social media messages crafted by PLM communications staff were brief headlines announcing the PVA project; they highlighted research in PD, Dr. Max Little, Sage Bionetworks, and participation in voice research. Additionally, Dr. Max Little also participated in social media outreach, messaging his Twitter followers and communicating on Facebook.

All recruitment messaging was submitted and approved by WIRB prior to its being used. In total, the PVA landing page received 5151 hits during the 8-week period of data collection in

January and February 2014. This contributed to the finally tally of 818 PDRS-reference number pairs (after data cleaning) from over 620 unique individuals. Some individuals captured their voice recording and PDRS multiple times over the course of the data collection.

Data storage in Synapse

The data is stored on the Sage bionetworks Synapse analysis platform. Synapse enables researchers to seamlessly and transparently conduct, track and share their ongoing work – building up living research projects in real-time. The platform consists of a web portal, web services, integrations with data analysis tools and is organized around novel "Analytical Communities" that scientists can create or join to work on complex medical problems together. Use of this platform for this patient voice analysis challenge allows analysts to work together and build off of one another's analytical designs in order to develop the best possible model of PD severity from voice recordings. Researchers can access the project (Synapse ID: 2321745 https://www.synapse.org/#!Synapse:syn2321745), document their work with the project wiki functionality, submit their analysis results and custom algorithms and ask questions, share ideas, and report problems.

Data access for this analysis was limited to invited analysts who are registered on Synapse and who contractually agree to (1) the Synapse Terms and Conditions of Use which contain multiple statements regarding protection of human data, the agreement not to try to identify or contact human subjects whose data is being analyzed, to do no harm, and not to redistribute the data; to (2) working collaboratively with proper attributions and acknowledgement of all involved and; (3) to limiting use of the data to within the confines of this pilot Challenge and keeping the data confidential and secure.

Data Overview and Analysis

Of the 818 PDRS collected, 779 were matched to voice recordings. After removing samples that failed the voice feature extraction procedure, we retained 747 usable PDRS/voice pairs. The available covariates include:

(i) Years since first symptom
(ii) Participant's current age
(iii) Days from recording to PDRS
(iv) On/off treatment indicator
(v) Participant's sex

A complete data description is provided in the separate stand-alone analysis report provided in Appendix.

In our baseline data analysis, we considered responses to the FullPDRS, MotorPDRS, and NonMotorPDRS. The FullPDRS score corresponds to the sum of the scores across all 17 questions in the questionnaire, re-scaled from [0-68] to range between 0 and 100 as described in the analysis report. The NonMotorPDRS score corresponds to the re-scaled sum of scores of questions 1 to 4 of the questionnaire (concerning memory problems, hallucinations/visions, mood, and motivation). The MotorPDRS score corresponds to the re-scaled sum of the scores of the 13 remaining questions (concerning speech, excessive saliva, swallowing, handwriting, cutting food, dressing, hygiene, turning in bed, falling, freezing, walking, tremors, numbness).

In this benchmark analysis we generated 81 distinct models using 9 alternative predictive methods and 9 distinct data type combinations. Predictive models included: (1) baseline model; (2) linear regression; (3) best subset selection in linear regression; (4) ridge-regression; (5) lasso; (6) elastic-net; (7) k-nearest neighbors regression; (8) random forests; and (9) boosted regression trees. The data combinations included the FullPDRS, MotorPDRS, and NonMotorPDRS and the 3 distinct types of feature data (namely, Voice plus Covariate features, Voice features alone, and Covariate features alone).
We also investigated the predictive performance of ensemble predictors derived from these 9 models using stacked regression. Our findings suggest that more stringent quality control is needed in order to improve predictive performance of the voice features obtained in uncontrolled settings. The detailed results are provided in a separate document added in Appendix.

Challenge data analysis Approach

We initiated a closed analysis challenge to test whether novel analytical approaches could be used to build models of disease severity from these data. Selected experts were invited to collaboratively map the PDRS data against voice characteristics, and to validate the ability to predict specific elements of impact /disability of disease like tremor or slowness. One goal for this closed challenge was to determine whether a larger effort is desirable.

Invited dry-run participants were graduate students from Stanford big data mining class as well as analysts from the University of Bochum (Germany), Israel Institute of technology (Israel), Universiidad Politecnica de Madrid (Spain), University of Oxford (UK), Northeastern University (US). Each analyst registered an account on Synapse to access the dataset. To conduct the challenge we randomly split the dataset into a training dataset and a test dataset of 389 and 390 rows respectively. The entire training set and test set inputs (i.e., the voice features and covariates) were available to the dry run participants, whereas the test set outputs (responses) were hidden during the model building phase, and were available only at the final model scoring phase.

We engage participants to ask questions and share expertise via email and scheduled webinar.

The R code used in the benchmark analyses is available in Github:
https://github.com/echaibub/PredictiveModelingPipeline

Issues and lessons learned

Lessons learned - Outreach and recruitment

The outreach effort required to reach people to participate in a fairly complex user experience in order to volunteer data was intense, and the need to bring in skilled communications experts early to craft and direct the placement of messages to individuals and groups was clear. Coordinating timing of press and media efforts within the limited timeframe of this seedling project was a challenge; had the media coverage occurred earlier in the data collection time frame, more people may have participated. A key lesson learned was that patients, when approached in a coordinated fashion with clear instructions, will participate in a complex data entry exercise. Future similar work would benefit from development of a complete communications campaign in the earliest planning stages of the project as a whole, and

include thinking through the types of information that could be fed back to participants at the conclusion of the project as well.

Lessons learned - PLM Data collection

Participants who provided feedback during and after the collection of data pointed to several areas which could be improved in future work:

- Better error messages with instructions for patients in cases of incorrect reference number entry
- Better visualization of correct entry for patients' confirmation
- Clearer guidelines for multiple entries by a single individual

From a technical standpoint, synchronizing the server times between the voice and PLM data servers would have made the process of matching the PDRS done in conjunction with the voice recording much easier. Asking patients in the instructions to minimize background noise may have helped the data quality (although perhaps not by as much as other issues identified in the voice data lessons learned below). Finally, an issue that was recognized early was that PD can create cognitive issues with patients, making the fairly complex set of tasks necessary to complete the data collection even more difficult for more severely affected PD patients. Whether this could be remedied by technical simplification of the process or would require the assistance of another person to guide them through the steps is an open question for future work.

Lessons learned- Voice data

Analysis of features extracted from the voice recordings which passed the rudimentary QC algorithm were not predictive; that is, they were most likely not useful for predicting the clinical outcomes in this setup. Extracting meaningful data from voice recordings collected in the wild proved challenging due to variations in interpretation of instructions by participants and by variations in the sound quality as a result of degradation in the signal transmission path. Manual spot checks show that many voice recordings are corrupted by severe transmission path problems, for example, aggressive automatic gain control or analog line noise. An enhanced QC pipeline should help detect the combined ambient background noise and occasional sound artifacts present in many voice samples.

Lessons learned - Dry-run challenge

Frequent communication facilitates the exchange of ideas and allows junior researchers to feel more involved in the project. Considering the small number of selected analysts and their diverse expertise, we needed to use diverse approaches to facilitate engagement. Webinars worked well but dealing with time zone differences complicated scheduling. A web forum could help document questions and support the nascent community.

Beyond the Pilot phase- Future directions and potential applications

Clearly one of the most important issues to address will be to develop effective voice recording QC editing algorithm that have low type I/II error rates. This will require novel research because little effort has been put into this area. We believe a promising direction would be in the area of semi-supervised learning approaches

Beyond the Pilot phase - The DREAM Challenge approach

We believe that the best approach toward developing robust and accurate predictive models of disease is to enable an open diverse community where data access is simple and people are incentivized to share their analysis methods and results. From previous experience, we learned that prize-based data analysis challenges and/or competitions are very effective at soliciting contributions from skilled analysts. This is especially true for data analysis problems that can be solved using sophisticated machine learning and data mining methods. The main advantage of open challenges lies in encouraging a diversity of analytical approaches from across scientific disciplines to solve inherently difficult but important questions. Last year, Sage Bionetworks used this approach to successfully identify genomic biomarkers predictive of breast cancer survival [14] and to demonstrate that bioinformatics specialists were better able to build a successful predictive biomarker of breast cancer survival when working together as a community in comparison to working alone.

A future open challenge would be designed to analyze various recordings with the goal to (1) find voice biomarkers for high frequency disease monitoring and (2) accurately assess disease severity and progression.

In summary, we believe that this pilot study successfully demonstrated the difficulties in acquiring high quality data from uncontrolled and unedited audio sources, and in making predictions based on such data. This pilot helped us to gain insight into what is necessary to conduct a larger-scale study for the assessment of disease severity and progression of Parkinson's disease based on voice data.

Acknowledgments

Many people contributed to this project. We particularly thank the participants who agreed to share their responses to PDRS and voice recordings.

References

1. Goetz, C. G. *et al.* Movement Disorder Society Task Force report on the Hoehn and Yahr staging scale: status and recommendations. *Mov. Disord.* **19,** 1020–1028 (2004).
2. Willis, A. W., Schootman, M., Evanoff, B. A., Perlmutter, J. S. & Racette, B. A. Neurologist care in Parkinson disease: A utilization, outcomes, and survival study. *Neurology* **77,** 851–857 (2011).
3. The Unified Parkinson's Disease Rating Scale (UPDRS): status and recommendations. *Mov. Disord.* **18,** 738–50 (2003).
4. Goetz, C. G. *et al.* Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Mov. Disord.* **23,** 2129–70 (2008).
5. Marras, C. *et al.* Survival in Parkinson disease: thirteen-year follow-up of the DATATOP cohort. *Neurology* **64,** 87–93 (2005).
6. Tsanas, A., Little, M. A., McSharry, P. E. & Ramig, L. O. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE Trans. Biomed. Eng.* **57,** 884–893 (2010).
7. Dorsey, E. R. *et al.* Increasing access to specialty care: a pilot, randomized controlled trial of telemedicine for Parkinson's disease. *Mov. Disord.* **25,** 1652–1659 (2010).
8. Venkataraman, V., Donohue, S. J., Biglan, K. M., Wicks, P. & Dorsey, E. R. Virtual visits for Parkinson disease: A case series. *Neurol Clin Pr.* 01.CPJ.0000437937.63347.5a–(2013). doi:10.1212/01.CPJ.0000437937.63347.5a
9. Wicks, P. & MacPhee, G. J. A. Pathological gambling amongst Parkinson's disease and ALS patients in an online community (PatientsLikeMe.com). *Mov. Disord.* **24,** 1085–1088 (2009).
10. Little, M., Wicks, P., Vaughan, T. & Pentland, A. Quantifying short-term dynamics of Parkinson's disease using self-reported symptom data from an Internet social network. *J. Med. Internet Res.* **15,** e20 (2013).
11. Wadman, M. New cures sought from old drugs. *Nature* **490,** 15 (2012).
12. Plenge, R. M. *et al.* Crowdsourcing genetic prediction of clinical utility in the Rheumatoid Arthritis Responder Challenge. *Nat. Genet.* **45,** 468–9 (2013).
13. Norman, T. C., Bountra, C., Edwards, A. M., Yamamoto, K. R. & Friend, S. H. Leveraging crowdsourcing to facilitate the discovery of new medicines. *Sci. Transl. Med.* **3,** 88mr1 (2011).
14. Margolin, A. A. *et al.* Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci. Transl. Med.* **5,** 181re1 (2013).
15. Derry, J. M. J. *et al.* Developing predictive molecular maps of human disease through community-based modeling. *Nat. Genet.* **44,** 127–130 (2012).
16. Pateints Like Me. Openness Philosophy. at <http://www.patientslikeme.com/about/openness>
17. Kinnunen, T. & LI, H. An overview of text-independent speaker recognition: from features to supervectors. *Speech Commun.* (2010). at <http://cs.joensuu.fi/sipu/pub/SRE-review-Kinnunen-and-Li.pdf>

Appendix

PVA Challenge Dry Run Report

by Elias Chaibub Neto and the PVA Challenge Dry Run Team

1 Introduction

The Patient Voice Analysis project (PVA Challenge) is an effort to assess the severity and fluctuations in Parkinson's disease symptoms using voice recordings. The goals of the project include:

1. Prove that crowdsourcing approaches to collect internet-based voice recordings can be linked to self-reported outcomes and analyzed safely in a distributed competition to validate the use of health sensors in diagnostic and biomedical research.

2. Combine three methods that have each been established independently;

   (a) Collecting validated patient-reported outcomes (PRO) through an online patient community (PatientsLikeMe).

   (b) Mass collection of voice samples to characterize Parkinsons disease (Parkinsons Voice Initiative).

   (c) Community-based, reproducible collaborative data analysis on Synapse (Sage Bionetworks).

3. Identify barriers and solutions to expand the scope of this work into a novel, rapid, cost-effective way of validating the therapeutic value of new health sensors in brain diseases.

This report focus on item 2(c), more specifically on the results of a dry-run exercise to evaluate the viability of the present data for a crowdsourcing machine learning prediction challenge.

2 Data overview

The PVA data set consists of: (i) brief voice recordings of sustained voice phonations (3-30 seconds long); (ii) self-reported symptom assessment (PDRS - Parkinson's Disease Rating Scale as well as Hoehn & Yahr stage); and (iii) a limited set of covariates from 620 individuals with Parkinson's disease (although the data set contains 779 samples, since some individuals captured their voice recording and PDRS multiple times over the course of the data collection).

The PDRS questionnaire data was summarized into 3 distinct response variables, namely, FullPDRS, MotorPDRS, and NonMotorPDRS. The FullPDRS score corresponds to the sum of the scores across all 17 questions in the questionaire, re-scaled to range between 0 and 100 as described by transformation in eq. (2) of the Methods section. The NonMotorPDRS score corresponds to the re-scaled sum of scores of questions 1 to 4 of the questionaire (concerning memory problems, hallucinations/visions, mood, and motivation). The MotorPDRS score corresponds to the re-scaled sum of the scores of the 13 remaining questions (concerning speech, excessive

saliva, swallowing, handwriting, cutting food, dressing, hygiene, turning in bed, falling, freezing, walking, tremors, numbness).

The available covariates include: (i) years since first symptom; (ii) participant's current age; (iii) days from pvi to pdrs; (iv) on treatment indicator; and (v) participant's sex. After processing, 38 voice features were extracted from the raw voice recording data (see Section 5.2 in Methods for further details). After removing samples that failed the voice feature extraction procedure, we ended up with 747 samples. Figures 1, 2, and 3 represent, respectively, the response, covariate, and voice feature distributions across the 747 samples.

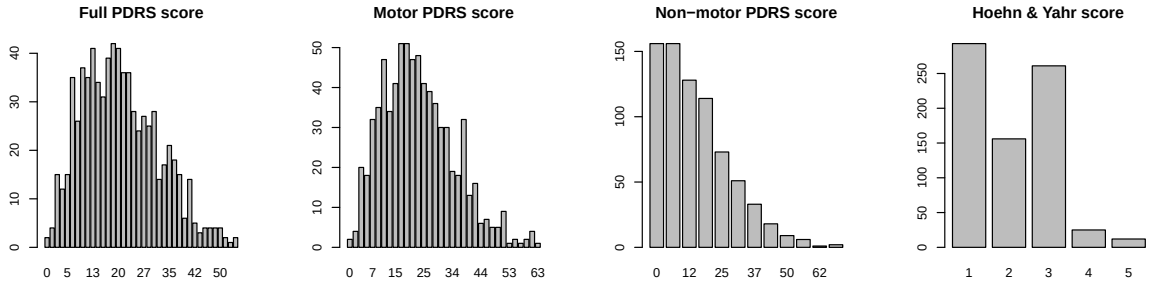


Figure 1. Response variable distributions.

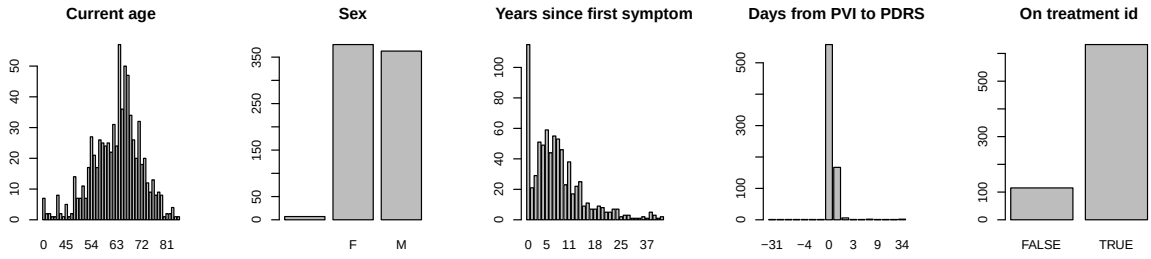


Figure 2. Covariate distributions. Summary statistics for the current age covariate are: Min. is 32; 1st Qu. is 58; Median is 64; 3rd Qu. is 68; Max. is 92. The sex covariate has a rather balanced distribution (note as well that 7 participants have missing sex data). Summary statistics for years since first symptom covariate are: Min. is 0 ; 1st Qu. is 3; Median is 7; 3rd Qu. is 11; Max. is 49. The days from PVI to PDRS covariate represents the number of days between the PVI voice recording and answering the PDRS questionnaire. The on treatment id covariate indicates whether the participant's answers to the questionnaire represent the participant's health state when going through medical treatment for Parkinson's disease.

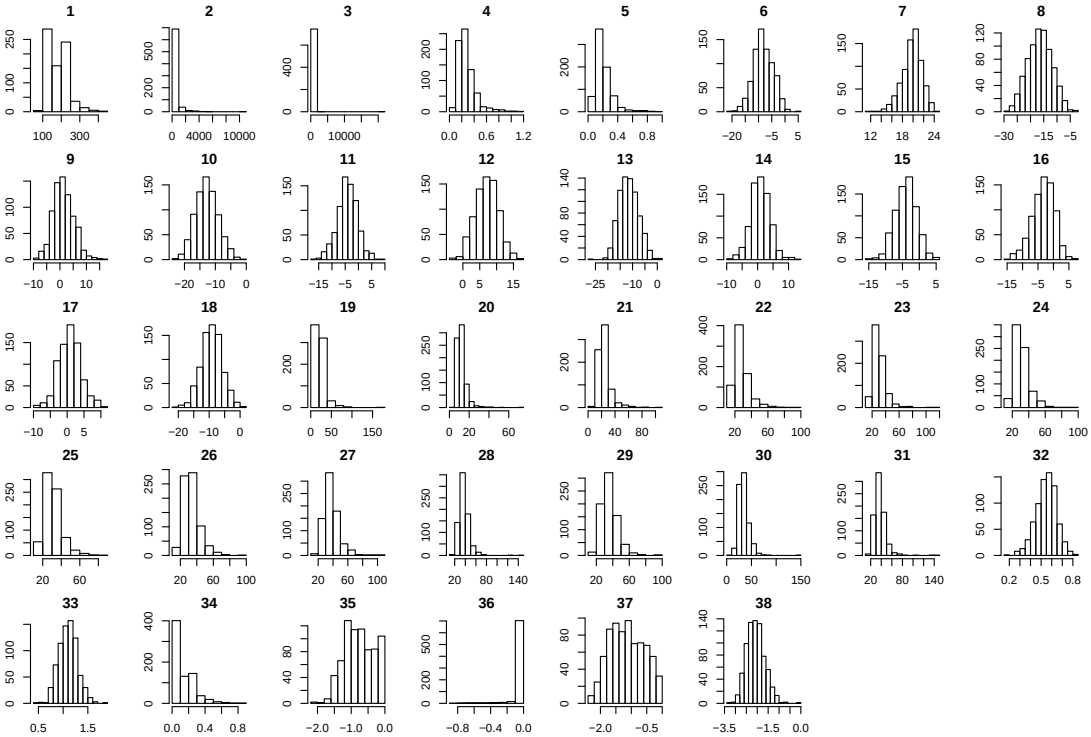


Figure 3. Distributions of the 38 voice features. Note the long tails for some of the features.

Figure 4 presents the distribution of the number of calls per patient across the 747 samples. Note that some of the participants made several calls. In order to account for the possible bias introduced by this issue, we assess predictive performance using Weighted Absolute Mean Error (WMAE) defined in eq. 1 on Section 5.3. The idea is to assign a weight proportional to the number of calls made by the participant during the computation of the error measure, so that participants that called multiple times are penalized to a greater extent.

2.1 Training/test data split

After initial processing (where we discarded the quality control variables present data files), the full data set was composed of 779 rows (samples) and 47 columns (4 responses, 5 covariates, and 38 voice features). The full data was randomly split into a training and a test set, with the training and test sets composed, respectively, of 389 and 390 rows and 47 columns. The entire training set and test set inputs (i.e., the voice features and covariates) were available to the dry run participants, whereas the test set outputs (responses) were hidden during the model building phase, and were available only at the final model scoring phase.

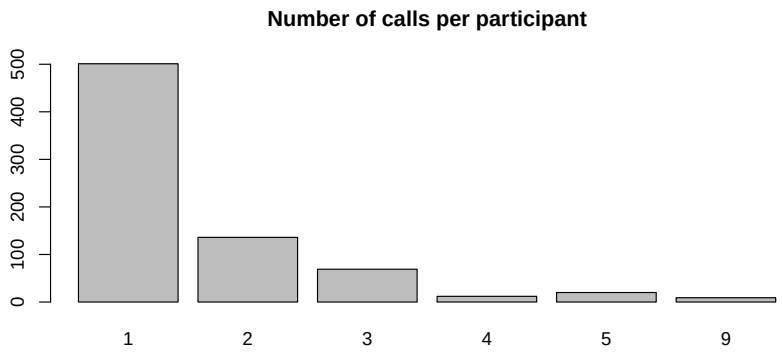


Figure 4. Distribution of the number of call per participant. Most of the participants made one, two, or three calls (to a lesser extent), while a few of them made as much as nine calls.

2.1.1 Data processing prior to predictive analysis

Prior to the analysis we removed samples which failed the voice features extraction process (32 samples), as well as samples from covariates with unusual values, such as: days from pvi to pdrs different from 0 or 1; sex different from M or F; and current age equal to 0. (Representing 30 additional samples in total.) After processing we are left with 363 samples in the training set, and 354 in the test set. (Although, for models generated using only the voice features as inputs, we have 374 training and 373 test samples, since we do not need to remove samples with unusual covariate values in this case.)

3 Results

In this study we considered a total of 81 distinct models generated from the combination of 9 alternative predictive methods and 9 distinct data type combinations. Predictive models included : (1) baseline model; (2) linear regression; (3) best subset selection in linear regression; (4) ridge-regression; (5) lasso; (6) elastic-net; (7) k-nearest neighbors regression; (8) random forests; and (9) boosted regression trees. Brief descriptions of each of these models are given in the Methods section.

Data combinations included 3 distinct versions of PDRS scores as response variables (namely, FullPDRS, MotorPDRS, and NonMotorPDRS) and 3 distinct types of feature data (namely, Voice plus Covariate features, Voice features alone, and Covariate features alone). Explicitly we considered the 9 distinct response/feature data combinations:

1. FullPDRS $\sim$ Voice + Covariates ,
2. FullPDRS $\sim$ Voice ,
3. FullPDRS $\sim$ Covariates ,

4. MotorPDRS $\sim$ Voice + Covariates ,
5. MotorPDRS $\sim$ Voice ,
6. MotorPDRS $\sim$ Covariates ,
7. NonMotorPDRS $\sim$ Voice + Covariates ,
8. NonMotorPDRS $\sim$ Voice ,
9. NonMotorPDRS $\sim$ Covariates .

We also investigated the use of the Hoehn & Yahr score as response variable in preliminary studies. However, since PDRS scores produced better predictive models (lower WMAEs) than the Hoehn & Yahr score we did not pursue a systematic evaluation of this response.

The on treatment id and sex variables were converted to 0 and 1 numerical variables for the analyses. Table 1 presents the results. The first column represents the average WMAE obtained from 100 random splits of the training data into train and test sets (see Section 4.3 for further details). The second column reports the standard deviations of the WMAE values across the 100 data splits. The third column shows the model ranks relative to the mean WMAE score. The fourth column shows the WMAE score generated from the test data, and the fifth column presents the model ranks relative to the test WMAE score.

Table 1

	mean WMAE	sd WMAE	train.rank	test.WMAE	test.rank
random_forest___FullPDRS~Covs	8.193980	0.5749795	7	8.013441	1
knn___FullPDRS~Covs	8.075521	0.5959760	4	8.038318	2
boosted_regr_trees___FullPDRS~Covs	7.908658	0.5873613	3	8.039886	3
random_forest___FullPDRS~Voice+Covs	7.835783	0.5455000	2	8.202845	4
ridge___FullPDRS~Covs	8.487461	0.5668257	17	8.306173	5
lasso___FullPDRS~Covs	8.462333	0.5689799	15	8.309387	6
enet___FullPDRS~Covs	8.460393	0.5748202	14	8.315277	7
lin_regr___FullPDRS~Covs	8.442973	0.5748839	13	8.315860	8
step_regr___FullPDRS~Covs	8.484009	0.5841078	16	8.355877	9
boosted_regr_trees___FullPDRS~Voice+Covs	7.729247	0.6155224	1	8.390096	10
boosted_regr_trees___MotorPDRS~Covs	8.221956	0.7095127	8	8.427596	11
enet___FullPDRS~Voice+Covs	8.118310	0.5656721	5	8.557384	12
random_forest___MotorPDRS~Covs	8.588902	0.6813709	19	8.559716	13
knn___MotorPDRS~Covs	8.511762	0.7265455	18	8.595023	14
lasso___FullPDRS~Voice+Covs	8.133043	0.5818538	6	8.607612	15
random_forest___MotorPDRS~Voice+Covs	8.274709	0.6756244	11	8.626649	16
ridge___MotorPDRS~Covs	8.891726	0.6747993	30	8.713349	17
lasso___MotorPDRS~Covs	8.892637	0.6727493	31	8.725929	18
baseline___FullPDRS~Voice	9.037896	0.6923016	38	8.732422	19
step_regr___FullPDRS~Voice+Covs	8.283595	0.6145271	12	8.745996	20
ridge___FullPDRS~Voice+Covs	8.258851	0.5902245	10	8.770021	21
enet___MotorPDRS~Covs	8.883878	0.6727333	29	8.794584	22
lin_regr___MotorPDRS~Covs	8.896184	0.6735661	33	8.799857	23
lin_regr___FullPDRS~Voice+Covs	8.918981	0.9530648	35	8.831481	24
knn___FullPDRS~Voice	9.021909	0.6929274	37	8.856996	25
boosted_regr_trees___MotorPDRS~Voice+Covs	8.253306	0.7000987	9	8.861740	26
step_regr___MotorPDRS~Covs	8.892789	0.6865982	32	8.884017	27
random_forest___FullPDRS~Voice	8.747120	0.6586102	23	8.887801	28
knn___FullPDRS~Voice+Covs	8.914689	0.5700772	34	8.890006	29
ridge___FullPDRS~Voice	8.805413	0.6701187	26	8.901193	30

enet___FullPDRS~Voice	8.777653	0.6768349	25	8.923405	31
boosted_regr_trees___FullPDRS~Voice	8.706214	0.6625214	22	8.942533	32
baseline___FullPDRS~Voice+Covs	9.078035	0.5942962	39	8.958989	33
baseline___FullPDRS~Covs	9.078035	0.5942962	39	8.958989	33
lasso___FullPDRS~Voice	8.754554	0.6621017	24	8.973574	35
step_regr___FullPDRS~Voice	8.968508	0.7139686	36	9.028310	36
lasso___MotorPDRS~Voice+Covs	8.656469	0.6924544	20	9.093403	37
lin_regr___FullPDRS~Voice	9.374992	0.8780526	42	9.129465	38
enet___MotorPDRS~Voice+Covs	8.660581	0.6644558	21	9.130731	39
ridge___MotorPDRS~Voice+Covs	8.839492	0.7270601	27	9.194688	40
lin_regr___MotorPDRS~Voice+Covs	9.624063	1.0335438	50	9.446512	41
knn___MotorPDRS~Voice+Covs	9.581627	0.6819399	47	9.461568	42
knn___MotorPDRS~Voice	9.666653	0.7825283	53	9.461769	43
baseline___MotorPDRS~Voice	9.617087	0.7702778	49	9.495513	44
step_regr___MotorPDRS~Voice+Covs	8.840316	0.7043192	28	9.503766	45
ridge___MotorPDRS~Voice	9.463473	0.7617486	46	9.609614	46
baseline___MotorPDRS~Voice+Covs	9.663216	0.7103101	51	9.619011	47
baseline___MotorPDRS~Covs	9.663216	0.7103101	51	9.619011	47
lasso___MotorPDRS~Voice	9.402146	0.7474376	43	9.667200	49
random_forest___MotorPDRS~Voice	9.442266	0.7264635	45	9.713765	50
boosted_regr_trees___MotorPDRS~Voice	9.326775	0.7306328	41	9.716532	51
enet___MotorPDRS~Voice	9.416441	0.7556730	44	9.725982	52
lin_regr___MotorPDRS~Voice	10.225932	0.9923258	54	9.915196	53
step_regr___MotorPDRS~Voice	9.608288	0.8357558	48	9.945366	54
knn___NonMotor~Covs	10.491322	0.8290812	62	10.029184	55
random_forest___NonMotor~Covs	10.441672	0.8070462	59	10.053463	56
random_forest___NonMotorPDRS~Voice+Covs	10.360133	0.7826068	57	10.059660	57
knn___NonMotorPDRS~Voice+Covs	10.731724	0.8891549	74	10.091452	58
enet___NonMotor~Voice	10.470464	0.8711339	61	10.128553	59
baseline___NonMotor~Voice	10.764301	0.8577279	78	10.220173	60
lasso___NonMotor~Voice	10.536195	0.8727843	66	10.226482	61
ridge___NonMotor~Voice	10.506341	0.8675245	63	10.235363	62
boosted_regr_trees___NonMotor~Covs	10.643928	0.7847657	73	10.248729	63
knn___NonMotor~Voice	10.806006	0.8860769	79	10.251954	64
boosted_regr_trees___NonMotorPDRS~Voice+Covs	10.310482	0.7700057	55	10.266945	65
enet___NonMotorPDRS~Voice+Covs	10.338556	0.8217419	56	10.273948	66
random_forest___NonMotor~Voice	10.579819	0.9142684	69	10.305554	67
step_regr___NonMotor~Covs	10.737424	0.8402194	75	10.311378	68
step_regr___NonMotorPDRS~Voice+Covs	10.551599	0.8641324	68	10.395844	69
ridge___NonMotorPDRS~Voice+Covs	10.412068	0.7944695	58	10.401358	70
lin_regr___NonMotor~Voice	11.263128	1.0836740	81	10.425308	71
lasso___NonMotorPDRS~Voice+Covs	10.459881	0.8082003	60	10.429212	72
boosted_regr_trees___NonMotor~Voice	10.508723	0.8798534	64	10.437330	73
enet___NonMotor~Covs	10.545641	0.8155637	67	10.482174	74
baseline___NonMotorPDRS~Voice+Covs	10.756304	0.8144344	76	10.488312	75
baseline___NonMotor~Covs	10.756304	0.8144344	76	10.488312	75
lin_regr___NonMotor~Covs	10.529311	0.8228276	65	10.489993	77
step_regr___NonMotor~Voice	10.642743	0.8521140	72	10.574906	78
ridge___NonMotor~Covs	10.619366	0.8025912	70	10.643536	79
lasso___NonMotor~Covs	10.637442	0.8055809	71	10.743625	80
lin_regr___NonMotorPDRS~Voice+Covs	11.078091	1.2043574	80	10.784431	81

Table 1 shows some interesting patterns. First, there is a strong correlation (0.93) between the model ranks from the training data alone (third column of Table 1) and the model ranks obtained from the test data WMAE scores (last column of Table 1). Figure 5 illustrates this strong rank agreement.

Second, it is easier to predict the FullPDRS response (note that the top 10 models were generated with the FullPDRS), whereas the NonMotorPDRS response is clearly harder to predict (models generated with the NonMotorPDRS performed worse than models generated with the

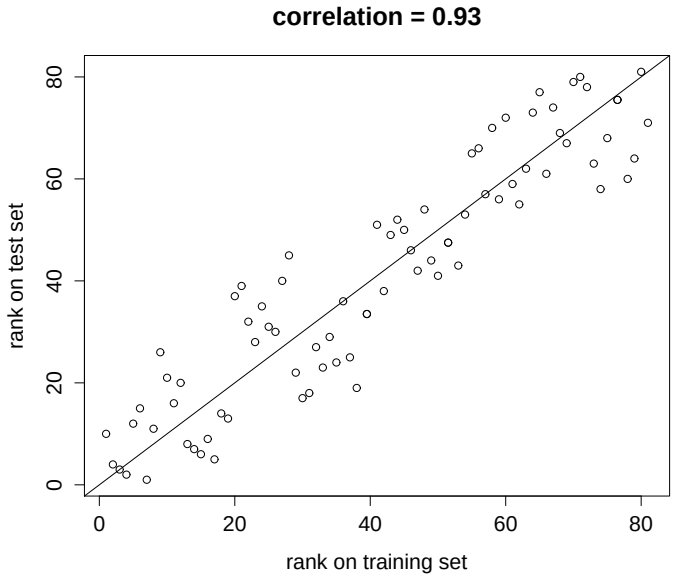


Figure 5. Ranking agreement between WMAE scores estimated from the training data alone and the real test WMAE scores. This suggests there were no serious over-fitting problems during the model training phase.

FullPDRS and MotorPDRS, ranking from position 55 to 81). The MotorPDRS response showed intermediate performance.

Third, we observed that only 2 out of the top 10 ranking models used voice features. The best performing model including voice features ranked in the fourth position and corresponds to a random forest. The absolute best model was (again) a random forest, but only employing the covariates data. Tables 2 and 3 present, respectively, the rankings of all 9 predictive methods for data combinations FullPDRS $\sim$ Voice + Covariates and FullPDRS $\sim$ Covariates. Figure 6 shows the importance plots for these top performing random forest models in their respective data combinations. The panels clearly show that the years since first symptom covariate has the strongest contribution to FullPDRS prediction.

Table 2

FullPDRS ~ Voice + Covs

	mean WMAE	sd WMAE	train.rank	test.WMAE	test.rank
random_forest	7.835783	0.5455000	2	8.202845	1
boosted_regr_trees	7.729247	0.6155224	1	8.390096	2
enet	8.118310	0.5656721	3	8.557384	3
lasso	8.133043	0.5818538	4	8.607612	4
step_regr	8.283595	0.6145271	6	8.745996	5
ridge	8.258851	0.5902245	5	8.770021	6
lin_regr	8.918981	0.9530648	8	8.831481	7
knn	8.914689	0.5700772	7	8.890006	8
baseline	9.078035	0.5942962	9	8.958989	9

Table 3

FullPDRS ~ Covs

	mean WMAE	sd WMAE	train.rank	test.WMAE	test.rank
random_forest	8.193980	0.5749795	3	8.013441	1
knn	8.075521	0.5959760	2	8.038318	2
boosted_regr_trees	7.908658	0.5873613	1	8.039886	3
ridge	8.487461	0.5668257	8	8.306173	4
lasso	8.462333	0.5689799	6	8.309387	5
enet	8.460393	0.5748202	5	8.315277	6
lin_regr	8.442973	0.5748839	4	8.315860	7
step_regr	8.484009	0.5841078	7	8.355877	8
baseline	9.078035	0.5942962	9	8.958989	9

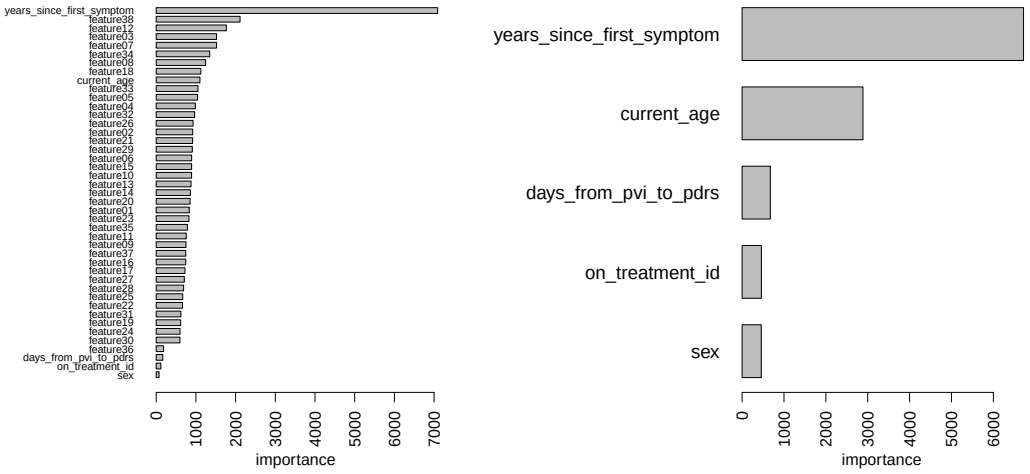


Figure 6. The left panel shows the importance plot for the random forest model fitted to data combination FullPDRS $\sim$ Voice + Covariates. The right panel shows the importance plot for the random forest model fitted to data combination FullPDRS $\sim$ Covariates.

Finally we point out that, in addition to the 9 predictive models employed in this analysis, we also investigated the predictive performance of ensemble predictors derived from these 9 models using stacked regression. We tried 4 distinct stacked regression approaches, namely, stacking using non-negative weights (Breiman's original proposal), as well as stacking with ridge-, elastic-net, and lasso-regression. Due to the amount of time needed to generated the cross-validated predictions to be used as features in the stacking approaches we did not apply stacking the 100 random split analyses. Table 4 reports test WMAE scores for all 81 original models plus the respective stacked regressions.

Table 4

test WMAE

stacked_regr_lasso___FullPDRS~Covs	7.914344
stacked_regr_ridge___FullPDRS~Covs	7.917397
stacked_regr_enet___FullPDRS~Covs	7.919509
stacked_regr_nnls___FullPDRS~Covs	8.003286
random_forest___FullPDRS~Covs	8.013441
knn___FullPDRS~Covs	8.038318
boosted_regr_trees___FullPDRS~Covs	8.039886
stacked_regr_enet___FullPDRS~Voice+Covs	8.186172
random_forest___FullPDRS~Voice+Covs	8.202845
stacked_regr_lasso___FullPDRS~Voice+Covs	8.203628
stacked_regr_ridge___FullPDRS~Voice+Covs	8.269964
ridge___FullPDRS~Covs	8.306173
lasso___FullPDRS~Covs	8.309387
enet___FullPDRS~Covs	8.315277
lin_regr___FullPDRS~Covs	8.315860
stacked_regr_nnls___FullPDRS~Voice+Covs	8.340410
step_regr___FullPDRS~Covs	8.355877
boosted_regr_trees___FullPDRS~Voice+Covs	8.390096
boosted_regr_trees___MotorPDRS~Covs	8.427596
stacked_regr_ridge___MotorPDRS~Covs	8.430687
stacked_regr_enet___MotorPDRS~Covs	8.444317
stacked_regr_lasso___MotorPDRS~Covs	8.453826
stacked_regr_nnls___MotorPDRS~Covs	8.467721
enet___FullPDRS~Voice+Covs	8.557384
random_forest___MotorPDRS~Covs	8.559716
knn___MotorPDRS~Covs	8.595023
lasso___FullPDRS~Voice+Covs	8.607612
random_forest___MotorPDRS~Voice+Covs	8.626649
stacked_regr_lasso___MotorPDRS~Voice+Covs	8.685205
stacked_regr_enet___MotorPDRS~Voice+Covs	8.693974
stacked_regr_ridge___MotorPDRS~Voice+Covs	8.706566
ridge___MotorPDRS~Covs	8.713349
lasso___MotorPDRS~Covs	8.725929
stacked_regr_nnls___MotorPDRS~Voice+Covs	8.726012
baseline___FullPDRS~Voice	8.732422
stacked_regr_enet___FullPDRS~Voice	8.736363
stacked_regr_lasso___FullPDRS~Voice	8.739872
step_regr___FullPDRS~Voice+Covs	8.745996
stacked_regr_ridge___FullPDRS~Voice	8.768562
ridge___FullPDRS~Voice+Covs	8.770021
enet___MotorPDRS~Covs	8.794584
stacked_regr_nnls___FullPDRS~Voice	8.797032
lin_regr___MotorPDRS~Covs	8.799857
lin_regr___FullPDRS~Voice+Covs	8.831481
knn___FullPDRS~Voice	8.856996
boosted_regr_trees___MotorPDRS~Voice+Covs	8.861740
step_regr___MotorPDRS~Covs	8.884017
random_forest___FullPDRS~Voice	8.887801
knn___FullPDRS~Voice+Covs	8.890006
ridge___FullPDRS~Voice	8.901193
enet___FullPDRS~Voice	8.923405
boosted_regr_trees___FullPDRS~Voice	8.942533
baseline___FullPDRS~Voice+Covs	8.958989
baseline___FullPDRS~Covs	8.958989
lasso___FullPDRS~Voice	8.973574
step_regr___FullPDRS~Voice	9.028310
lasso___MotorPDRS~Voice+Covs	9.093403
lin_regr___FullPDRS~Voice	9.129465
enet___MotorPDRS~Voice+Covs	9.130731
ridge___MotorPDRS~Voice+Covs	9.194688
lin_regr___MotorPDRS~Voice+Covs	9.446512
knn___MotorPDRS~Voice+Covs	9.461568

knn___MotorPDRS~Voice	9.461769
baseline___MotorPDRS~Voice	9.495513
step_regr___MotorPDRS~Voice+Covs	9.503766
stacked_regr_ridge___MotorPDRS~Voice	9.561119
stacked_regr_lasso___MotorPDRS~Voice	9.569926
stacked_regr_enet___MotorPDRS~Voice	9.590778
stacked_regr_nnls___MotorPDRS~Voice	9.607050
ridge___MotorPDRS~Voice	9.609614
baseline___MotorPDRS~Voice+Covs	9.619011
baseline___MotorPDRS~Covs	9.619011
lasso___MotorPDRS~Voice	9.667200
random_forest___MotorPDRS~Voice	9.713765
boosted_regr_trees___MotorPDRS~Voice	9.716532
enet___MotorPDRS~Voice	9.725982
lin_regr___MotorPDRS~Voice	9.915196
step_regr___MotorPDRS~Voice	9.945366
stacked_regr_nnls___NonMotor~Covs	10.029088
knn___NonMotor~Covs	10.029184
stacked_regr_nnls___NonMotorPDRS~Voice+Covs	10.044700
random_forest___NonMotor~Covs	10.053463
stacked_regr_enet___NonMotorPDRS~Voice+Covs	10.057514
random_forest___NonMotorPDRS~Voice+Covs	10.059660
stacked_regr_ridge___NonMotorPDRS~Voice+Covs	10.065154
stacked_regr_lasso___NonMotor~Voice	10.067721
stacked_regr_enet___NonMotor~Voice	10.068828
stacked_regr_nnls___NonMotor~Voice	10.073586
stacked_regr_ridge___NonMotor~Voice	10.075618
stacked_regr_lasso___NonMotorPDRS~Voice+Covs	10.082959
knn___NonMotorPDRS~Voice+Covs	10.091452
stacked_regr_lasso___NonMotor~Covs	10.097851
stacked_regr_ridge___NonMotor~Covs	10.119828
enet___NonMotor~Voice	10.128553
stacked_regr_enet___NonMotor~Covs	10.132179
baseline___NonMotor~Voice	10.220173
lasso___NonMotor~Voice	10.226482
ridge___NonMotor~Voice	10.235363
boosted_regr_trees___NonMotor~Covs	10.248729
knn___NonMotor~Voice	10.251954
boosted_regr_trees___NonMotorPDRS~Voice+Covs	10.266945
enet___NonMotorPDRS~Voice+Covs	10.273948
random_forest___NonMotor~Voice	10.305554
step_regr___NonMotor~Covs	10.311378
step_regr___NonMotorPDRS~Voice+Covs	10.395844
ridge___NonMotorPDRS~Voice+Covs	10.401358
lin_regr___NonMotor~Voice	10.425308
lasso___NonMotorPDRS~Voice+Covs	10.429212
boosted_regr_trees___NonMotor~Voice	10.437330
enet___NonMotor~Covs	10.482174
baseline___NonMotorPDRS~Voice+Covs	10.488312
baseline___NonMotor~Covs	10.488312
lin_regr___NonMotor~Covs	10.489993
step_regr___NonMotor~Voice	10.574906
ridge___NonMotor~Covs	10.643536
lasso___NonMotor~Covs	10.743625
lin_regr___NonMotorPDRS~Voice+Covs	10.784431

As expected, stacking improved predictive performance (note that the 4 top ranked models were stacked models). However, once again, these best ranking stacked models used only the covariate data.

4 Discussion

Given the noisy nature of the voice data (recall that the voice recordings were made in an uncontrolled setting) and the tight time frame available for performing the voice feature extraction, it is not surprising that the voice features were not very predictive of the PDRS scores. Furthermore, it is important to point out that telephone-based voice collection is a pioneering enterprize and as yet, quality control protocols specifically tailored to uncontrolled settings are still under active development. For instance, after observing the present analysis results (and obtaining similar results in independent analyses) our collaborator, Max Little, went back to the raw voice data and detected a fair amount of flawed samples which, nevertheless, had passed the quality control protocol developed for the analysis of laboratory controlled recordings. These findings suggest that more stringent quality control is needed in order to improve predictive performance of the voice extracted features in uncontrolled settings.

5 Methods

5.1 PDRS questionnaire data

The PDRS questionnaire is an abbreviated version of the Unified Parkinson's Disease Rating Scale (UPDRS) widely used to evaluate Parkinson's Disease severity. The 17 items PDRS questionnaire omits the clinical observation section present in the more comprehensive 42-item UPDRS. PDRS can be self-administered and completed quickly (approximately 10 minutes). PDRS has a maximum score of 68 points. Each question is rated on a (0-4) scale with "0" representing no disability and "4" worst disability. The PDRS score is the "floor" of the sum of the scores of the 17 items rescaled from [0-68] to [0-100] as described in detail in Section 5.4. In addition to the 17 PDRS questions, the PDRS questionnaire includes the Hoehn & Yahr Staging (1-5).

5.2 Voice data

In addition to the questionnaire items, the data also includes 38 features extracted from the voice recordings. We refer to Synapse's wiki,

`https://www.synapse.org/#!Synapse:syn2321745/wiki/62077`

for a description of the extracted voice features.

5.3 Predictive performance evaluation metric

The full data was randomly split into a training and a test set (of roughly equal size), and predictive performance of competing methods was evaluated using the weighted mean absolute error (WMAE) metric defined as,

$$\text{WMAE} = \sum_{i=1}^{N} w_i |y_i - \hat{y}_i| \ , \tag{1}$$

where N represents the test set size; y_i represents a the test set outcome; $\hat{y}_i$ represents the prediction for outcome y_i; and w_i represents the weight of prediction $\hat{y}_i$, defined as $w_i = N_i/N$, where N_i represents the number of times subject i appears in the test set. (Note that this measure penalizes to a greater extent the subjects that made multiple calls.)

5.4 Response transformation

The PDRS scores were computed as follows. For any number of questions q, let s represent the sum of the q questions scores. Since each question can be scored as 0, 1, 2, 3, or 4, we have that s ranges from 0 to $4q$. The PDRS is them computed as

$$\text{PDRS} = \left\lfloor \frac{100\, s}{4\, q} \right\rfloor , \tag{2}$$

where the floor function $\lfloor x \rfloor$ represents the largest integer not greater than x. Note that this transformation makes the PDRS score range between 0 and 100, independent of the number q of questions used in the computation of the score. When comparing different PDRS scores (i.e., FullPDRS, MotorPDRS, and NonMotorPDRS) it is important to have all scores in the same range so that the WMAE derived from models using different responses can still be compared.

5.5 Model ranking during the training phase

During the training phase (where we did not have assess to the test data) we evaluated and ranked the predictive performance of the different models investigated in this study by performing 100 random splits of the training data into training and test sets (with each random split dividing the training data into 5 roughly equal sized parts, with 4 parts used for training, and the remaining one for testing), and computing the average WMAE score across the 100 random splits of the training data. All models were trained and tested in the same 100 random data splits.

5.6 Tuning parameter optimization

For all methods depending on tuning parameters, we performed tuning parameter optimization via 10 fold cross-validation. WMAE as adopted as the predictive performance criterium during the cross-validation process.

5.7 Models

5.7.1 Baseline model

The baseline model corresponds to the average outcome prediction, $\hat{y}_i = \bar{y}_{train}$, $i = 1, \ldots, N_{test}$, where N_{test} represents the number of test samples and $\bar{y}_{train}$ corresponds to the average outcome in the training set. There are no tuning parameters in this inflexible approach.

5.7.2 Linear regression

Multiple linear regression fit was done using the standard `lm` function in R base installation. There are no tuning parameters.

5.7.3 Best subset selection in linear regression

We fitted best subset variable selection for linear regression using the `step` R function (in R base installation) with BIC penalty and search direction set to "both" (i.e., performing both forward selection and backward variable elimination). There are no tuning parameters.

5.7.4 Penalized linear regression: ridge, lasso, and elastic-net regression

Lasso, elastic-net, and ridge-regression were fitted using the `glmnet` package in R. For all 3 models, we adopted the default λ tuning parameter grid generated automatically (in a data-driven fashion) by the software. For elastic-net we adopted the grid $\{0, 0.1, 0.2, \ldots, 1\}$ for the α tuning parameter, and cross-validated both λ and α parameters.

5.7.5 K-nearest neighbors regression

Knn-regression was fitted with the `knn.reg` function from the `FNN` package in R. We adopted the grid $\{1, 2, 3, \ldots, 30\}$ for the number of neighbors tuning parameter.

5.7.6 Random forests

Random forest regression was fitted using the default options of the `randomForest` function from the `randomForest` R package.

5.7.7 Boosted regression trees

Boosted regression trees were fitted with the `gbm` function of the `gbm` R package. We set `distribution` equal to "gaussian" and used number of trees (`n.trees`) as a tuning parameter (while adopting the default values of all other parameters).

5.7.8 Stacked regression

The basic idea in stacked-regression is to produce an ensemble model generated as a linear combination of cross-validated predictors from a series of models. A cross-validated predictor is produced by splitting the training data into f folds, training a model in $f - 1$ folds, and predicting the left out fold (the cross-validated predictor corresponds to the catenation of all f predictions generated in this way). We consider 4 distinct versions of stacked-regression: (i) stacked-regression using non-negative least squares (Breiman's original proposal); (ii) stacked-regression using ridge-regression; (iii) stacked-regression using lasso; and (iv) stacked-regression using elastic-net.

6 Code

The R code used in the analyses is available in Github:

https://github.com/echaibub/PredictiveModelingPipeline