Journal of Parkinson's Disease 13 (2023) 757–767
DOI 10.3233/JPD-225083
IOS Press

Clinical Research

What Patients Say: Large-Scale Analyses of Replies to the Parkinson's Disease Patient Report of Problems (PD-PROP)

Connie Marras[a,*], Lakshmi Arbatti[b], Abhishek Hosamath[b], Amy Amara[c], Karen E. Anderson[d], Lana M. Chahine[e], Shirley Eberly[f], Dan Kinel[g], Sneha Mantri[h], Soania Mathur[i], David Oakes[f], Jennifer L. Purks[g], David G. Standaert[j], Caroline M. Tanner[k], Daniel Weintraub[l] and Ira Shoulson[b,g]
[a]*Edmond J Safra Program in Parkinson's Disease, University Health Network, University of Toronto, Toronto, Canada*
[b]*Grey Matter Technologies, a Wholly Owned Subsidiary of Modality.ai, San Francisco, CA, USA*
[c]*Department of Neurology, University of Colorado Anschutz Medical Campus, Aurora, CO, USA*
[d]*Departments of Psychiatry and Neurology, Georgetown University, Washington DC, USA*
[e]*Department of Neurology, University of Pittsburgh, Pittsburgh, PA, USA*
[f]*Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, USA*
[g]*Department of Neurology, University of Rochester, Rochester NY, USA*
[h]*Department of Neurology, Duke University, Durham, NC, USA*
[i]*PD Avengers, Toronto, Canada*
[j]*Department of Neurology, University of Alabama at Birmingham, Birmingham, AL, USA*
[k]*Department of Neurology, Weill Institute for Neurosciences, University of California – San Francisco, San Francisco, CA, USA*
[l]*Departments of Psychiatry and Neurology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA*

Accepted 14 May 2023
Pre-press 9 June 2023
Published 25 July 2023

Abstract.
Background: Free-text, verbatim replies in the words of people with Parkinson's disease (PD) have the potential to provide unvarnished information about their feelings and experiences. Challenges of processing such data on a large scale are a barrier to analyzing verbatim data collection in large cohorts.
Objective: To develop a method for curating responses from the Parkinson's Disease Patient Report of Problems (PD-PROP), open-ended questions that asks people with PD to report their most bothersome problems and associated functional consequences.
Methods: Human curation, natural language processing, and machine learning were used to develop an algorithm to convert verbatim responses to classified symptoms. Nine curators including clinicians, people with PD, and a non-clinician PD expert classified a sample of responses as reporting each symptom or not. Responses to the PD-PROP were collected within the Fox Insight cohort study.

*Correspondence to: Connie Marras, MD, PhD, Toronto Western Hospital, 399 Bathurst St, 7 McL, Toronto, ON M5T 2S8, Canada. Tel.: +1 416 603 6422; Fax: +1 416 603 5004; E-mail: connie.marras@uhnresearch.ca.

Results: Approximately 3,500 PD-PROP responses were curated by a human team. Subsequently, approximately 1,500 responses were used in the validation phase; median age of respondents was 67 years, 55% were men and median years since PD diagnosis was 3 years. 168,260 verbatim responses were classified by machine. Accuracy of machine classification was 95% on a held-out test set. 65 symptoms were grouped into 14 domains. The most frequently reported symptoms at first report were tremor (by 46% of respondents), gait and balance problems (>39%), and pain/discomfort (33%).
Conclusion: A human-in-the-loop method of curation provides both accuracy and efficiency, permitting a clinically useful analysis of large datasets of verbatim reports about the problems that bother PD patients.

Keywords: Patient-reported outcome, measurement, Parkinson's disease, machine learning

INTRODUCTION

Increasingly, there is an emphasis on patient-reported outcomes (PROs) for use in clinical research [1]. The vast majority of such instruments restrict answers to a pre-specified range of responses. While they allow the respondent to report on their health state, the possible set of responses is constrained to lie within the preconceived structure of the scale. Measures that allow the respondent to report their health state without such restrictions are rare and have been minimally incorporated into quantitative clinical research, either interventional or observational. The advantages of such instruments include capturing a fuller range of patient experience and potentially increasing sensitivity to unanticipated effects of disease or interventions. On the other hand, instruments allowing open-ended responses are challenging to handle from privacy, data management, and analysis perspectives, again restricting the scale on which such data can be collected.

The Parkinson's Disease Patient Reports of Problems (PD-PROP) is a series of open-ended questions that asks people with Parkinson's disease (PD) to report and rank, in their own words, up to five PD-related bothersome problems and their functional consequences, without restriction of content or length. The PD-PROP has been incorporated as a module within Fox Insight [2], an online, observational, longitudinal clinical study that has collected, as of February 2022, anonymous participant-reported outcomes on approximately 53,000 individuals with and without PD. At such a scale, machine-assisted solutions are necessary and, as shown by related experience, cannot be independent of human oversight in order to ensure that the output is interpretable and clinically relevant [3, 4]. To address this challenge, we have developed a data curation approach that combines humans (clinicians, other subject matter experts and experience experts (people with PD)) with natural language processing and machine learning to create a dataset that captures the spectrum and frequency of symptoms that matter most to patients. This process is an expansion of an initial curation of the PD-PROP data introduced previously [5]. Herein we describe in detail the expanded curation process and its results.

MATERIALS AND METHODS

The PD-PROP

The PD-PROP is completed by participants enrolled in Fox Insight (https://foxinsight.michaeljfox.org), an online, observational, longitudinal clinical study sponsored by the Michael J. Fox Foundation for Parkinson's Research (MJFF) that anonymously collects self-reported health related information, as well as lifestyle information and previous exposures [2]. Data used in the preparation of this article were obtained from the Fox Insight database (https://foxinsight-info.michaeljfox.org/insight/explore/insight.jsp) on 03/02/2020. For up-to-date information on the study, visit https://foxinsight-info.michaeljfox.org/insight/explore/insight.jsp.

PD-PROP comprises open-ended questions that ask people with PD to report, in their own words, up to five PD-related bothersome problems and their functional consequences. The two questions to be answered for each problem reported are: 1) What is the most bothersome problem for you due to your Parkinson's disease? 2) In what way does this problem bother you by affecting your everyday functioning or ability to accomplish what needs to be done? See Supplementary Table 1 for the full instrument. Participants respond on-line by keyboard entry. Each problem and its associated consequence as entered by the participant are combined and referred to as a 'verbatim". Participants are invited to complete the PD-PROP at 3-month intervals.

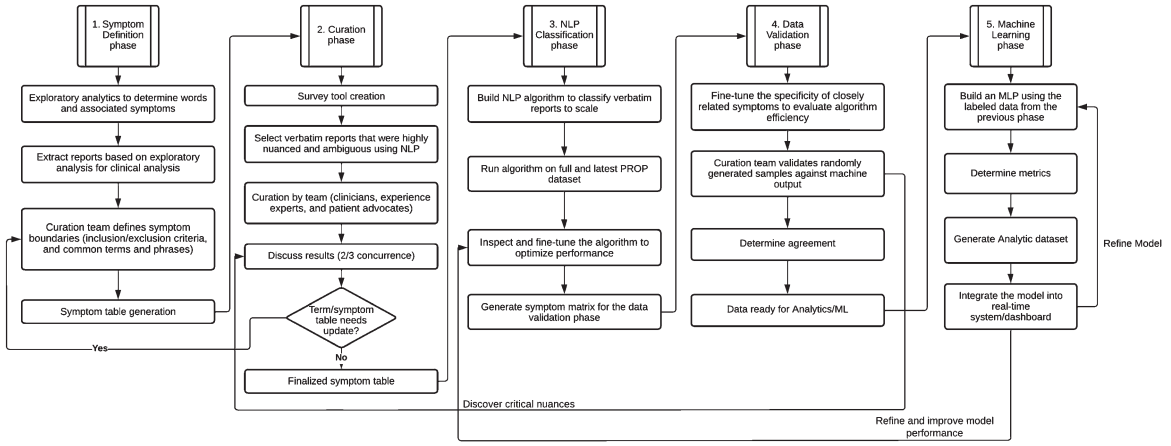

Fig. 1. Human-in-the-Loop Curation and Classification Methodology. NLP, natural language processing; PROP, patient report of problems.

Human-in-the-loop curation and classification algorithm development

Curation is the process of identifying symptoms from the problems expressed in the verbatims and classifying each verbatim as mentioning or not mentioning a specific symptom. Our method involved human curators who provided classification for a sample of verbatims. This experience provided terms and phrases that informed the development of a natural language processing (NLP) algorithm to classify verbatims at scale. The entire process is depicted in Fig. 1 and described in detail below.

Symptom definition phase

As an initial triaging and exploratory approach, data were visualized using a combination of latent Dirichlet allocation topic modelling [6] and generating uni-gram and bi-gram word clouds [7, 8]. This approach proved helpful to initiate a list of topically identified symptoms for further granularization. From then on, based on 1) knowledge of the motor and non-motor symptoms of PD, 2) review of over 5300 PD-PROP responses in prior curation work [5, 9], and 3) review of approximately 25 sample verbatims by each curation team member tailored to their specific area of expertise, the curation team generated a list of symptoms that were anticipated to be mentioned in the larger dataset of PD-PROP responses, and grouped them by domain (e.g., "cognition" or "autonomic"). The team of 9 curators consisted of clinicians, people with PD, and a non-clinician PD expert with extensive experience talking to people with PD in her work for a Parkinson's support and research charity. The clinician members have a broad range of training and relevant expertise related to PD, including motor aspects, cognition and psychiatry, narrative medicine, outcomes research, sleep medicine, and family practice. The relevant subject matter experts developed boundaries for each problem, stipulating inclusion and exclusion criteria, conceptually. The language used to define boundaries avoids medical terms, when possible, to reflect the patient voice and to facilitate use by non-medical curation team members and future data users. Proposed boundaries were circulated among the curators and refined until consensus was reached.

Curation and symptom term table generation

During the discussions establishing problem boundaries, words and phrases that were likely to represent each symptom were proposed and formed the basis for initial sampling of the database for the first round of curation. A sample of 50 verbatims was extracted from the February 2020 Fox Insight PD-PROP dataset for each symptom using the proposed words and phrases. Verbatims that were highly nuanced or potentially ambiguous were preferentially selected. In addition, up to four verbatims were repeated in each set to assess intra-rater reliability. A group of three curators, consisting of at least one person with PD or the non-clinician PD expert and at least one clinician, was assigned to each symptom. Symptoms were assigned to groups based on the expertise of the clinician member. Group members independently classified each verbatim as reporting or not reporting the particular symptom as defined by the boundaries and reported the specific terms and

phrases that were key to their decision. They were instructed to consider each problem referred to by the respondent without considering attribution or severity. For example, if fatigue was reported and attributed to insomnia in the verbatim, that verbatim would be classified as reporting both fatigue and insomnia as a bothersome problem, not only the underlying insomnia thought responsible for the fatigue. This was to reduce bias and preserve the meaning of the responses.

As curators reviewed verbatims they were also asked to identify symptoms, using their clinical knowledge and personal experience, that were not represented in the problem list that had been developed. After the initial round of classification, each curation group met in breakout sessions to review examples of disagreement in order to build consensus on classification and to adjust symptom names. When it was evident that two or more symptoms were difficult to differentiate in the verbatims, a more general symptom was created. Conversely, when one symptom was found to encompass several distinct concepts reflected in verbatims, the symptom was split into two or more symptoms. This process resulted in an initial algorithm that incorporated the curator-provided list of relevant words and phrases. The algorithm was further expanded using synonym generation through Unified Medical Language System (UMLS) ontologies [10, 11] and NLP techniques such as word vectorization [12–14], which provided closely related terms and phrases that had the same intended meaning [15]. A finalized symptom term table was then created which would serve as the input to the next phase.

These determinations and associated terms and phrases served to develop the machine learning algorithm, which utilized a combination of human-in-the-loop and standard natural language processing and data analytics techniques [16, 17].

Natural language processing (NLP) classification [16, 17]

The data were first cleaned to identify spelling errors using a Java language-based package [18, 19]. The advantage of using this package over Peter Norvig's algorithm-based Python autocorrect module [19] was that some of the specific jargon such as cramping and migraine that were incorrectly autocorrected to "tramping" and "migrate" by the Python algorithm were correctly retained by the Java module. An additional advantage was that we could also instruct the algorithm to ignore specific words from our library to prevent incorrect auto correction. The building of "ignore-terms" library is an ongoing process.

A database comprised of the user id of the participants and the verbatim (conjugation of problem and consequence) was created using Neo4j [20]. A set of external files comprised of terms and phrases provided by the curators along with their synonyms was used to perform phrase query extraction [21] on the database resulting in a master dataset that was comprised of each verbatim and its associated symptom binning. The results were then fine-tuned through manual inspection and the algorithm optimized. The dataset upon which the analysis was performed was then generated for data validation and generation of a scalable machine learning model as described in the sections that follow.

Data validation and optimization

From the full dataset of approximately 168,260 verbatims, between 455 and 600 verbatims were provided to each curator using the dataset resulting from the NLP classification phase. Each sample consisted of 11 or 12 positives, predicted by the algorithm to include the symptom of interest, and 12 negatives predicted not to include the symptom of interest. Negatives were enriched with examples from closely related symptoms in order to challenge the ability of curators to distinguish similar symptoms from each other (e.g., the sleep curation sample was enriched with verbatims predicted to report fatigue). Blind to this classification, each curator classified the provided verbatims as reporting or not reporting the symptom. The final expert consensus classification for a verbatim was designated as the determination provided by at least 2/3 curators in a group and was considered the 'gold standard' classification. Algorithm performance metrics were based on the concordance at this stage, and calculated as accuracy, proportion of false positives, and proportion of false negatives. Instances of discordance were discussed with curators at the discretion of the data science team, in order to optimize the algorithm, post-validation. Reasons for discordance informed modification of terms and phrases to identify positives and negatives.

Machine learning

Classification of verbatims into multiple symptoms is a multi-label text classification problem [22, 23].

We used a Keras-tensorflow [24] supervised deep learning neural network model to be trained on this data. The model comprised two hidden layers and one output layer. 10,519 unique multi-label combinations were identified in the dataset. We decided to use a 90-10 train-test split to ensure that we had sufficient data for training as well as testing, leaving us with ∼16,000 samples for testing. In order to ensure that data was not overfitted and to avoid any class imbalance problems, the test data were further split into test and validation sets in the ratio 1:1 providing us with ∼8000 test and ∼8000 validation samples. Sklearn's train-test-split was used to split the data [25]. The model was compiled using binary cross loss entropy.

RESULTS

Symptom boundaries

Consensus on an initial set of boundaries was reached on each definition after no more than two iterations. These were refined throughout the curation process to best reflect the self-reported experience of people with PD. For illustrative purposes, Table 1 shows the symptom list and associated boundaries that were developed for the cognition domain. The full list consists of 65 symptoms grouped into the following 14 domains: tremor, rigidity, bradykinesia, postural instability, gait, other motor, sleep, fatigue, cognition, affect/motivation/thought-perception/other psychiatric, pain, autonomic dysfunction, fluctuations, and dyskinesias and is available as Supplementary Table 2.

Curation and symptom table generation

A total of approximately 3500 verbatims in groups of approximately 50 per symptom were reviewed in the initial algorithm development phase. The process of curation was conducted over one year, consisting of approximately two hours of independent work between monthly meetings. From the initial round of review, complete (3/3) concordance across curators ranged from 28% of verbatims (for sudden OFF) to 100% (for sexual dysfunction), with median 3/3 concordance of 64%. At least 2/3 concordance ranged from 70% (for early morning awakening) to 100% (for 27 symptoms).

Terms and phrases for algorithm development were refined and expanded during the review process. Table 2 shows a sample of the terms and phrases for the symptoms internal tremor and executive abilities/working memory.

During the curation and algorithm development process, additional symptoms were identified that were not included in the validation exercise and for which boundaries had yet to be developed. These include fear of future events, medication side effects, restlessness, loss of sense of taste/smell, reduced self-esteem/embarrassment, tingling, and drooling.

Algorithm validation

As of February 2020, approximately 25,000 research participants with PD had completed at least one PD-PROP assessment used in the algorithm validation phase. The median age at first response to the PD-PROP was 67 years, 55% were male, and median years since PD diagnosis was 3 years.

168,260 verbatims were classified. Approximately 1% of verbatims were not classified in this iteration. Reasons for non-classification included verbatims in languages other than English, being uninterpretable, or belonging to symptoms not yet defined (such as 'fear of'). In the algorithm validation phase, full (3/3) concordance between curators for verbatims ranged from a low of 88% (for personality and behavior changes not otherwise specified) to 100% (for 59 symptoms). Sixty-eight verbatims were presented twice to each curator for a total of 204 pairs of duplicate presentations to assess intra-rater reliability. Ratings for 192/204 (94%) agreed across the two presentations. There was no evident clustering of discordant responses within specific symptoms. Concordance between the machine and the curators for the algorithm validation phase was based on individual curator classification for a set of verbatims. Twelve machine-assigned 'negatives' and 11 or 12 machine-assigned 'positives' enriched for conceptually-related symptoms (see Supplementary Table 3 for the grouping of related symptoms) were presented to the curators blind to machine classification. Accuracy (proportion of verbatims correctly classified) was lowest (96%) for several cognitive (memory, cognitive slowing/mental fatigue, visuospatial abilities) and depressive symptoms, and was 100% for 58/65 symptoms. For those symptoms with agreement less than 100%, there were no false negatives and the number of false positive responses did not exceed 1 for any. Machine learning model performance from an optimal run of 50 epochs yielded an F1 score ($2 \times \text{precision} \times \text{recall})/(\text{precision} + \text{recall}$)

Table 1
Conceptual boundaries for cognitive symptoms

Domain	Proposed reported symptom	Conceptual boundaries	
		Includes	Excludes
Cognition	memory	Impairment of memory including difficulty remembering information; learning new information; orientation to time, place	the term "having to remember"
	concentration/attention	Difficulty concentrating or paying attention; sustaining focus	
	cognitive slowing	Slowing or impairment of mental processing. Includes difficulty keeping up with conversations, slowness to respond, mental fatigue	confusion, 'brain fog', mental sharpness;
	language/word finding	difficulty understanding conversation; expressing oneself; difficulty speaking words that are being thought of. difficulty understanding what is being read/reading	difficulty understanding due to hearing impairment
	mental alertness/awareness	fluctuating alertness; fluctuations in/variable attention; zoning out, brain fog, confused thoughts, reduced mental sharpness	cognitive/mental slowing
	visuospatial abilities	difficulty judging distances or depth; navigating 3-dimensional situations; orienting oneself in space; identifying visual and spatial relationships among objects; trouble navigating closed or indoor spaces that are familiar	Freezing (interruption of gait) in doorways or thresholds.
	executive abilities/working memory	difficulty planning or executing tasks; multi-tasking; switching from one cognitive task to another, trouble following directions or instructions; problem solving; decision making; sequencing; learning new skills	
	cognitive impairment NOS	Cognitive complaint not clearly fitting into another category. Could include confusion, muddled, mixed up	

Table 2
Example terms and phrases for one motor and one non-motor symptom

Symptom	Sample terms/phrases
Internal Tremor	"internal tremor", "shake inside", "vibration"
Executive abilities/ working memory	"impaired mental flexibility", "cognitive adapting", "can't multi-task", "trouble planning and organizing", "starting or completing a task", "being adaptable or flexible", "difficulty following instructions"

Table 3
Performance of the algorithm in the machine learning phase

	Held-out test set
Accuracy	95%
F1 score	95%
Precision	97%
Recall	93%

F1 score: 2(precision)(recall)/(precision+recall).

of 0.95. Accuracy, precision and recall were 0.95, 0.97 and 0.93, respectively. Table 3 presents the performance of the algorithm in the machine learning phase.

Figure 2 shows the frequency of reports of bothersome problems within each domain at participants' first report by gender and age. To calculate frequencies, the denominator is the total number of symptoms reported across the cohort. For each participant only the first report of each symptom was counted. However, all responses to the PD-PROP over time were included, and a participant may report multiple symptoms at each time point. A higher percentage of women reported problems in all domains. The proportional gender differences were largest in the psychological, pain, fluctuations and dyskinesia domains. Compared with respondents below age 60, a

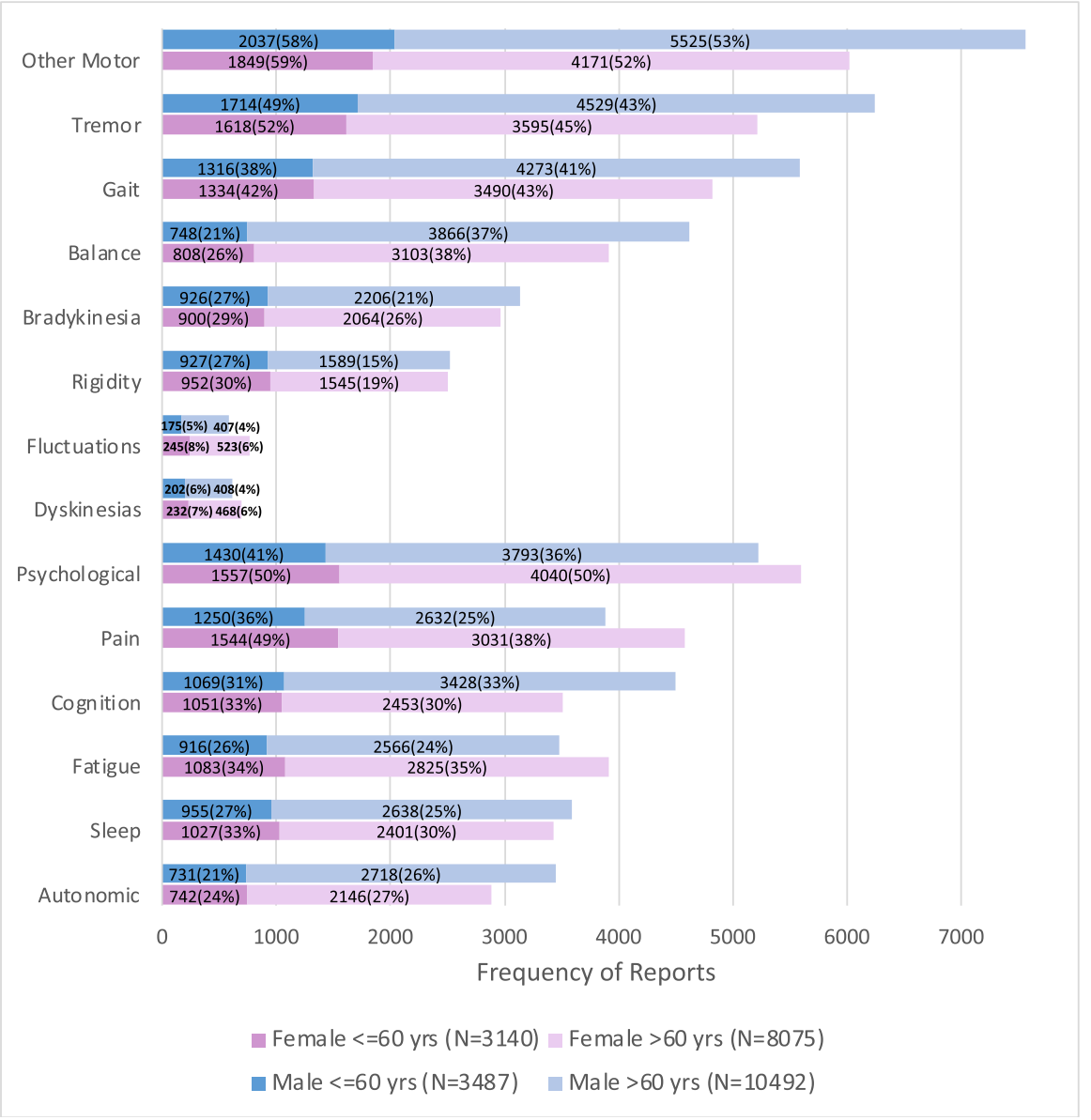

Fig. 2. Frequency* of patient reports of problems by domain, age, and gender. The denominator for percentages is the N given in the legend for the corresponding demographic category. *Calculated as the ratio of the number of individuals reporting the symptom at least once to the number of individuals in that demographic category.

higher proportion of those above age 60 reported gait, balance, and autonomic problems, cognitive problems (men only) and fatigue (women only).

The most frequently reported bothersome symptoms at first report were tremor (by 46% of respondents at first report), gait not otherwise specified (39%), and pain/discomfort (33%). Within the motor domains, the most commonly reported bothersome symptoms at first report were: tremor (46%), gait not otherwise specified (39%), impaired dexterity/micrographia (31%), and balance (29%). Fluctuations related to medication (3%) and dyskinesias (5%) were uncommonly reported as bothersome problems. Within the non-motor domains, a diverse array of symptoms was reported as being bothersome, the most common being pain (33%), physical fatigue (29%), and negative emotions not otherwise specified (22%). Table 4 shows the frequency of all symptoms that are reported by at least 1% of participants, categorized by domain, from most to least common.

Table 4
Curated Symptoms and their frequency at first PD-PROP report*

Domain	Symptom	Frequency (N)	% of all symptoms at first PD-PROP report
Tremor	Tremor	11454	46
	Internal Tremor	123	1
Rigidity	Stiffness	5013	20
Bradykinesia	Slowness	5913	24
	Facial Expression	270	1
Balance	Balance	7192	29
	Falls	2145	9
	Fear of Falling	1833	7
Gait	Gait NOS	9821	39
	Freezing of Gait	1366	5
Other Motor	Impaired Dexterity/Micrographia	7837	31
	Speech	5836	23
	Dystonia	2920	12
	Posture	982	4
Fluctuations	Off Periods - Medication Related	821	3
	Off Periods - Medication Not Mentioned	523	2
Dyskinesias	Dyskinesias	1310	5
Sleep	Poor Sleep Quality Unspecified	3528	14
	Sleep Maintenance Insomnia	1701	7
	Excessive Daytime Sleepiness	1559	6
	Sleep Onset Insomnia	487	2
	RLS/Restlessness	418	2
	RBD Like Symptoms	316	1
	Dreams	310	1
Fatigue	Physical Fatigue	7325	29
	Mental Fatigue	194	1
Cognition	Memory	3325	13
	Language/Word Finding	3132	12
	Concentration/Attention	2384	10
	Cognitive Slowing	1107	4
	Executive Abilities/Working Memory	828	3
	Mental Alertness/Awareness	732	3
	Cognitive Impairment NOS	302	1
Psychological	Negative Emotions or Cognition NOS	5611	22
	Anxiety/Worry	5005	20
	Depressive Symptoms	1827	7
	Apathy	963	4
	Loneliness/Isolation	540	2
	Hallucinations/Illusion/Presence/Passage	216	1
Pain	Pain/Discomfort	8389	33
	Cramp or Spasm	2736	11
Autonomic	Altered Bowel Frequency	2917	12
	Bladder Incontinence	1246	5
	Swallowing Problems	1036	4
	Lightheadedness/Dizziness	756	3
	Sexual Dysfunction	535	2
	Frequent Urination	452	2
	Nausea	259	1
	Bloating/Feeling Full	265	1
	Excessive Sweating	178	1

*Limited to symptoms having frequency of 1% or greater.

DISCUSSION

In the clinical setting patients report their problems and how these problems bother them. In turn, clinicians shape patient-reported problems and their functional consequences into defined symptoms (often elaborated by what makes the problems better or worse, and the severity or seriousness of the

problems for the patient). We have developed a data curation approach that is analogous to this process of clinical history-taking. Via PD-PROP, people with PD report problems which are classified into symptoms through an algorithm that combines human expert interpretation and machine learning. Incorporating human expert curation of the problems that people with PD report in their own words into the classification process provides clinical meaningfulness for classifying symptoms. Longitudinally, the resulting dataset and dictionary will inform a patient-reported natural history of PD [5].

A similar approach has been applied to people with PD in the United Kingdom, asking individuals to report up to three aspects of their condition they would like to see improved [26]. 790 participants provided responses, all of which were curated by a team of 6 people with direct experience with PD. Each response was categorized into one or more of 41 symptoms. This approach, relying solely on human curation of 100% of the submitted responses, has the advantage of a more thorough analysis of the responses, but is impractical for large scale data analysis or broadening the application of the instrument to other settings and over time. Using human-directed curation to inform the development of a machine algorithm permits scaling up to large datasets and replicable analyses over time.

To our knowledge, this combined human and machine method is novel, at least as applied in healthcare research. A recent systematic review of the use of machine learning methods to analyze patient-reported data did not identify examples of combined human and machine processes [27]. The identified studies used machine learning methods to analyze data from existing, traditional PRO measures with fixed-choice responses. The value of unconstrained response formats was highlighted in a study that examined the content of a "most important concern of the patient" question applied in routine clinical care of patients with multiple sclerosis. Responses included idiosyncratic symptoms, disease management, and social concerns not included in the established PRO measures, suggesting that some important concerns are not sufficiently captured by existing instruments [28]. Similarly, in the study of PD patients mentioned above [26], some patients prioritized issues such as better care and disease management, and the desire to maintain independence as those they most wanted to improve, demonstrating the importance of issues beyond symptoms in the experience of chronic disease.

We also encountered reported problems that were not anticipated and could not be mapped to our predetermined list of symptoms. Therein lies the value of free-text reporting, enriching the information collected by patient report beyond symptoms that are represented in commonly used PRO measures. An important example from our PD-PROP experience is the reporting of fear as a most bothersome problem. This experience mirrors a finding reported in a qualitative study asking people with chronic liver disease or renal transplants and researchers for feedback on existing PRO measures [29]. Patient participants and clinicians in that study suggested including a free text box where patients can mention other issues not covered in the questionnaire: "I think what's missing is the direct, what is the fear that you have of dying, what's the fear of you getting colon cancer, what's the fear of you getting bile duct cancer, and how does that fear manifest itself?" This fear symptom domain will be one of several additional problems curated in a next round of PD-PROP curation, examining a sample of responses expressing fears and defining categories of fears. Unanticipated symptoms that we incorporated into our curation included restlessness, mental fatigue, fear of falling, internal (inner) tremor, mental alertness/awareness, isolation, and nausea. The most frequent of these was fear of falling, reported in 7% of individuals at their first report. Other unanticipated symptoms yet to be curated include medication side effects, drooling, vertigo, and loss of sense of smell/taste.

A similar approach to collecting adverse event data from clinical trials supports the value of free text entry and demonstrates the feasibility of processing those data. Adverse event data was collected directly from participants in cancer clinical trials using electronic tablets at their study visits. Participants could choose from listed options or add free text. Physician researchers classified the free text-entered data. Of 1,357 free text entries, 87.5% were categorized as symptomatic adverse events, of which only 384 (32.4%) mapped to an existing National Cancer Institute library of terms used to classify adverse events in cancer clinical trials. These results suggested a number of additional terms to be added to the library, again demonstrating the importance of unconstrained responses [30].

Our curation experience revealed challenges mapping some reported problems onto symptoms, particularly in the cognition and sleep domains. This was often attributed to ambiguous responses, which in a clinical setting could be clarified with follow-up

questions. As a result, some symptoms that were proposed ahead of curation were merged due to inability to reliably interpret the entries to make a clear distinction between two concepts (for example, initially OFF periods was divided into sudden OFF, OFF periods medications not mentioned, and OFF periods medication-related). These were ultimately merged due to difficulty distinguishing them with confidence during curation. Despite such challenges, we were able to identify reported problems mapping to 65 different symptoms with high levels of agreement between curators.

The PD-PROP as applied in Fox Insight provides a curated dataset that demonstrates the frequency and broad spectrum of symptoms that matter most to people with PD and reflects the heterogeneity of PD. The large size of the dataset permits analysis within categories of age, sex, and disease duration and examination of their co-occurrence, which will further reveal the different phenotypes of PD. Ongoing longitudinal data collection will provide detailed insight into the variable progression of symptoms at an unprecedented scale. These findings can guide clinical care, research priorities, and provide guidance for clinical trial outcomes and sample size calculation [9]. By virtue of parallel collection of demographic, lifestyle, and clinical data in Fox Insight, there now exists a rich shared resource for evaluating associations between patient reported problems and quality of life, lifestyle factors, and other experiences of the person with PD (https://www.michaeljfox.org/fox-den). The longitudinal data collection allows researchers to investigate the predictive value of early symptoms for later outcomes, which can inform clinical trial design [9] as well as patient counseling.

Some limitations of the curation process deserve mention. First, because data entry was only possible by keyboard, individuals with dexterity difficulties may have been less likely to participate. The addition of a voice-entry option is being deployed to ease use and improve accessibility. Second, classification of the verbatim reports relies on interpretation by curators that could not be verified by the individual reporting the problem. Incorrect interpretations are possible, although the chances of this were reduced by a consensus approach to interpreting each verbatim response. Third, although the reports were generated by people with self-reported PD our curation was largely anchored in predetermined symptoms based on the clinical and personal experiences of the curators. Identifying unknown symptoms was dependent on *ad hoc* identification of reported problems that were not the target of the specific exercise. Therefore, our symptom list is biased toward already known symptoms with a chance of missing heretofore underrecognized symptoms. Conversely, the granularity of curation in some categories (e.g., different patterns of fluctuations of symptoms) resulted in small numbers of individuals being classified as reporting specific symptoms as most bothersome problems. Depending on the research aim, for the purpose of meaningful analysis it may be necessary to combine some symptoms together. Finally, despite the large number of responses reviewed, we did not identify all problems expressed. However, given the large number of verbatim responses reviewed it is unlikely that we missed frequently-reported problems.

In conclusion, a human-in-the-loop method of curation of patient-reported problems provides both accuracy and efficiency, permitting a clinically useful analysis of large datasets of free text responses. As longitudinal data accrues, we will learn about the natural history of PD as reported by people with PD directly and without constraint.

ACKNOWLEDGMENTS

The Fox Insight Study (FI) is funded by The Michael J. Fox Foundation for Parkinson's Research. We would like to thank the Parkinson's community for participating in this study to make this research possible.

CONFLICT OF INTEREST

Ira Shoulson, Lakshmi Arbatti, and Abhishek Hosamath are employees of Grey Matter Technologies, a wholly owned subsidiary of modality.ai.

DATA AVAILABILITY

The data supporting the findings of this study are not publicly available due to privacy restrictions. The curated datasets derived from the verbatim responses are publicly available in FoxDEN at https://foxden.michaeljfox.org upon signing a data use agreement.

SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: https://dx.doi.org/10.3233/JPD-225083.

REFERENCES

[1] Matts ST, Webber CM, Bocell FD, Caldwell B, Chen AL, Tarver ME (2022) Inclusion of patient-reported outcome instruments in US FDA medical device marketing authorizations. *J Patient Rep Outcomes* **6**, 38.

[2] Smolensky L, Amondikar N, Crawford K, Neu S, Kopil CM, Daeschler M, Riley L, Brown E, Toga AW, Tanner C (2020) Fox Insight collects online, longitudinal patient-reported outcomes and genetic data on Parkinson's disease. *Sci Data* **7**, 67.

[3] Binkheder S, Wu H-Y, Quinney SK, Zhang S, Zitu MM, Chiang CW, Wang L, Jones J, Li L (2022) PhenoDEF: A corpus for annotating sentences with information of phenotype definitions in biomedical literature. *J Biomed Semantics* **13**, 17.

[4] Nawab K, Ramsey G, Schreiber R (2020) Natural language processing to extract meaningful information from patient experience feedback. *Appl Clin Inform* **11**, 242-252.

[5] Javidnia M, Arbatti L, Hosamath A, Eberly SW, Oakes D, Shoulson I (2021) Predictive value of verbatim Parkinson's disease patient-reported symptoms of postural instability and falling. *J Parkinsons Dis* **11**, 1957-1964.

[6] Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J Mach Learn Res* **3**, 993-1022.

[7] Xu J, Tao Y, Lin H (2016) Semantic word cloud generation based on word embeddings. *IEEE Pacific Visualization Symposium (Pacific Vis)*, 239-243.

[8] Cui W, Wu Y, Liu S, Wei F, Zhou MX, Qu H (2010) Context preserving dynamic word cloud visualization. *IEEE Pacific Visualization Symposium (PacificVis)*, 121-128.

[9] Shoulson I, Arbatti L, Hosamath A, Eberly SW, Oakes D (2022) Longitudinal cohort study of verbatim-reported postural instability symptoms as outcomes for online Parkinson's disease trials. *J Parkinsons Dis* **12**, 1969-1978.

[10] Bodenreider O (2004) The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Res* **32**, D267-D270.

[11] National Library of Medicine, Unified Medical Language System, https://www.nlm.nih.gov/research/umls/index.html, Accessed January 11, 2023.

[12] Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. *arXiv*, arXiv:1301.3781.

[13] Chiu B, Crichton G, Korhonen A, Pyysalo S (2016) How to Train good Word Embeddings for Biomedical NLP. *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, Berlin, Germany, pp. 166-174.

[14] Chen Q, Peng Y, Lu Z (2019) BioSentVec: Creating sentence embeddings for biomedical texts. *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 1-5.

[15] Sabbir A, Jimeno-Yepes A, Kavuluru R (2017) Knowledge-based biomedical word sense disambiguation with neural concept embeddings. *Proc IEEE Int Symp Bioinformatics Bioeng* **2017**, 163-170.

[16] Zhang S, He L, Dragut E, Vucetic S (2019) How to invest my time: Lessons from human-in-the-loop entity extraction. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2305-2313.

[17] Wu X, Xiao L, Sun Y, Zhang J, Ma T, He L (2021) A survey of human-in-the-loop for machine learning. *ArXiv210800941 Cs*.

[18] Java Platform, Package java.lang, https://docs.oracle.com/javase/7/docs/api/java/lang/package-summary.html, Accessed February 24, 2023.

[19] Norvig P, How to Write a Spelling Corrector, https://norvig.com/spell-correct.html, February 24, 2023

[20] Neo4j, Neo4j Graph Database Platform. https://neo4j.com, Accessed May 6, 2020.

[21] Neo4j, Full-text search index - Cypher Manual, Neo4j Graph Data Platform.https://neo4j.com/docs/cypher-manual/5/indexes-for-full-text-search, February 28, 2023

[22] Keras, Keras documentation: Large-scale multi-label text classification. https://keras.io/examples/nlp/multi_label_classification, Accessed January 12, 2023.

[23] Hasan MM, Dip ST, Rahman T, Akter MS, Salehin I (2021) Multilabel movie genre classification from movie subtitle: Parameter optimized hybrid classifier. *2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*, pp. 1-6.

[24] Keras, Keras: The Python deep learning API, https://keras.io. February 28, 2023.

[25] scikit learn, sklearn.model_selection.train_test_split, scikit-learn, https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html, Accessed February 28, 2023.

[26] Port RJ, Rumsby M, Brown G, Harrison IF, Amjad A, Bale CJ (2021) People with Parkinson's disease: What symptoms do they most want to improve and how does this change with disease duration? *J Parkinsons Dis* **11**, 715-724.

[27] Verma D, Bach K, Mork PJ (2021) Application of machine learning methods on patient reported outcome measurements for predicting outcomes: A literature review. *Informatics* **8**, 56.

[28] Miller DM, Moss B, Rose S, Li H, Schindler D, Weber M, Planchon SM, Alberts J, Boissy A, Bermel R (2020) Obtaining patient priorities in a multiple sclerosis comprehensive care center: Beyond patient-reported outcomes. *J Patient Exp* **7**, 541-548.

[29] Aiyegbusi OL, Isa F, Kyte D, Pankhurst T, Kerecuk L, Ferguson J, Lipkin G, Calvert M (2020) Patient and clinician opinions of patient reported outcome measures (PROMs) in the management of patients with rare diseases: A qualitative study. *Health Qual Life Outcomes* **18**, 177.

[30] Chung AE, Shoenbill K, Mitchell SA, Dueck AC, Schrag D, Bruner DW, Minasian LM, St Germain D, O'Mara AM, Baumgartner P, Rogak LJ, Abernethy AP, Griffin AC, Basch EM (2019) Patient free text reporting of symptomatic adverse events in cancer clinical research using the National Cancer Institute's Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). *J Am Med Inform Assoc* **26**, 276-285.