# Deriving and Validating New "Outcome" Variables in Patient Reported Epidemiological Data

Rochelle E. Tractenberg[1,2,3], Futoshi Yumoto[2,3,4]
[1] Departments of Neurology; Biostatistics, Bioinformatics & Biomathematics; and Rehabilitation Medicine; Georgetown University, Washington, D.C.
[2] Collaborative for Research on Outcomes and –Metrics; Silver Spring, MD
[3] Psychometrics Core, Fox Insight Study (FI); The Michael J. Fox Foundation for Parkinson's Research
[4] Resonate, Inc., Reston, VA

**Abstract**
Patient reported epidemiological data are becoming more widely available for the application of data scientific methodologies and other investigations. One new such dataset, the Fox Insight project, was launched in 2017 by the Michael J. Fox Foundation to encourage researchers to engage in the study of Parkinson's disease. The Fox Insight (FI) dataset will be released for public access late in 2018, and while it is longitudinal in nature, and while dedicated statistics and psychometrics cores have supported the scientific advisory committee in their considerations of which variables to collect, diseases like Parkinson's do not have unambiguous states (e.g., "mild" vs. "severe") or changes (e.g., "stable" vs. "worse"). Assessing these states can be complicated when there are medical comorbidities that may contribute to, compound, or be conflated with the symptoms of the disease of interest (e.g., cerebrovascular disease; muscle weaknesses due to stroke or aging; depression). This paper describes the development and proposed validation of two new variables in the FI data set that are intended for use as "outcomes", which would be available for interested researchers who download the FI data when it becomes publicly available. One represents "cognitive change" and the other represents the "off" syndrome of Parkinson's where symptoms suddenly become unresponsive to medication (that works otherwise) for short periods of time. We discuss how new outcomes like these can be developed from patient reported epidemiologic data like the FI set using theory, and validated using international consensus criteria (COSMIN). These results are useful for planning analyses of the FI dataset, but also may support future designs for similar patient reported epidemiological data sets, so that they will be designed to include outcome variables or variables that can be used to demonstrate alignment of derived outcomes with the COSMIN validity criteria.[1]

**Key Words:** Patient Reported Outcomes, outcomes validation, COSMIN criteria, Parkinson's disease

1 Reprint requests: Rochelle Tractenberg, rochelle.tractenberg@gmail.com

## 1. Introduction

Parkinson's disease (PD) is a neurodegenerative disease that affects mobility and walking, balance, and coordination; other symptoms include cognitive decline, sleep disturbances, fatigue, and personality or behavioral changes. Although the pathology of PD is fairly well defined, in the sense that specific neurons in the substantia nigra area of the brain die off and lead to lower/decreasing levels of an important neurotransmitter, dopamine, diseases like PD do not actually have unambiguous states (e.g., "mild" vs. "severe"), nor are changes in symptoms (e.g., "stable" vs. "worse") well defined. Determining whether symptoms are emerging, worsening, or improving can be complicated when there are medical comorbidities that may contribute to, compound, or be conflated with the symptoms of the disease of interest, e.g., cerebrovascular disease; muscle weaknesses due to stroke or aging; depression – all of which can lead to sleep disturbances, fatigue, and cognitive, behavioral, and personality changes. Some cerebrovascular disease can also affect walking, balance and coordination.

PD is an area of intense research focus, with programs specifically funded by the US National Institutes of Health and other foundations.

In fact, in the United States, the National Institutes of Health has an ongoing policy encouraging/requiring the public sharing of data that are collected using federal resources (https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf). One new such dataset, the Fox Insight (FI) project, was launched by the Michael J. Fox Foundation in 2017 to encourage the study of Parkinson's disease and will be released for public access in 2019. Details about the study are given on the FI site (https://foxinsight.michaeljfox.org/) and the preliminary descriptive statistics have been discussed (cite paper).

To date (May 2018), 12,000 individuals with Parkinson's disease and 4,000 controls have contributed data to this resource. The resource is scheduled to be made public in Summer 2019. Work is ongoing to ensure that this dataset can function as a meaningful contribution to rigorous and reproducible science in Parkinson's disease (PD). Participants are contacted every 90 days and a series of surveys are administered, capturing patient reports on symptoms, activities of daily living, and other factors and demographic variables.

The Fox Insight (FI) dataset is longitudinal in nature, and dedicated statistics and psychometrics cores have supported the FI scientific advisory committee in their considerations of which variables to collect, but due to its fundamentally descriptive and patient reported nature, no "hard" outcome variables were included. The data could be mined using machine learning techniques, and while descriptive information would be obtained this way, associations among these variables is the strongest type of evidence that could be quantified.

Here we discuss the development, scoring considerations, and a validation approach for two new types of variables in the FI data set that could be used as "outcomes" for hypothesis-driven studies. These would be available for interested researchers who access the FI dataset when it becomes publicly available. One represents "cognitive change" and the other represents the "off" syndrome of Parkinson's where symptoms suddenly become unresponsive to medication (that works otherwise). We discuss how new outcomes like these can be developed from patient reported epidemiologic data like the

FI dataset using theory and clinical information, and validated using international consensus criteria (COSMIN, Mokkink et al. 2010).

## 2. Methods

These analyses proceeded by evaluating the content of all instruments included in the FI data set. The (current, May 2018) data dictionary was used to guide item identification, as described below. The FI data have not been analyzed here, only the *items in the dataset* that are relevant for the assessment of two key aspects of Parkinson's disease are analyzed for their scoring properties (Results): cognitive (non-motor) and OFF episode symptoms. The entire FI data dictionary was searched manually by the co-authors (who constitute the psychometrics core for this project).

Once the FI dataset is made public, scheduled for summer 2019, the data dictionary will also be made available. The scoring considerations described here will also be made available, supporting future research using the FI dataset from its date of release and beyond.

### 2.1 COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN)

According to the report of the international *Co*nsensus-based *S*tandards for the selection of health *M*easurement *In*struments (COSMIN, Mokkink et al. 2010), there are four psychometric aspects that are characterized as essential features of health related PRO by international agreement: *Reliability* (minimization of measurement error; internal consistency or interrelatedness of the items; and maximization of variability that is due to "true" difference between levels of the symptoms across patients), *validity* (*content*, reflection of the construct to be measured; *face*, recognizability of the contents as representing the construct to be measured; *structural*, the extent to which the instrument captures recognizable dimensions of the construct to be measured; and *criterion*, association with a gold standard), *cross-cultural validity*, and the *interpretability of scoring*. As has been argued elsewhere (Tractenberg et al., 2018), PRO data may not be specifically amenable to a structural validity analysis, particularly when it is highly patient-centered and/or has prioritized the patient perspective (rather than a more formal psychometric or research-driven perspective). Also, cross-cultural validity is less of a concern in these considerations because: a) while the FI data set does use some instruments that are in use worldwide, the cognitive and OFF items we examined here are not in use worldwide; and b) we assessed these two new outcomes from, and for, an existing dataset with a pre-defined cultural background (originating in the United States). Thus, we did not address either of these types of validity in our analyses.

Similarly, "responsiveness" is defined by the COSMIN report as reflecting the sensitivity of an instrument to change in the construct under study- and with the data set characteristics, and these variables selected out of a much larger set; besides which, the data are longitudinal but no actual endpoints or variables that could be considered endpoints or milestones are included. Thus, no gold standard for clinically meaningful change on any patient experience or other dimension of Parkinson's disease exists in the dataset. In fact, creating reliable and valid representations of clinically meaningful states and changes in those states is our objective in creating these new outcomes; these will need to be further validated in the future with independent samples where clinical outcomes *are* included.

The Results section first assesses the items in each of the outcome types (cognitive function/OFF episodes), commenting on scoring considerations; then the alignment of these outcomes with COSMIN criteria is described and discussed.

## 3. Results

### 3.1 "Cognitive status" and "cognitive change"

There are three sources for the assessment of cognitive status in the FI dataset. These sources are discussed below, followed by a section addressing the estimation of "cognitive change".

*3.1.1 Consistency in symptoms relating to cognitive status or function in the FI dataset*

Two of the three sources for the assessment of cognitive status ask for ratings of the frequencies of experiencing "symptoms" with respect to the past month. The first source includes the items in an "activities of daily living" inventory:

*i.* Due to having Parkinson's disease, how often during the last month have you had difficulty concentrating, e.g. when reading or watching TV?

*ii*. Due to having Parkinson's disease, how often during the last month have you felt unable to communicate with people properly?

The other source includes three items in the assessment of "non-movement" symptoms of Parkinson's disease in the last month:

*iii*. Have you experienced problems remembering things that have happened recently or forgetting to do things in the last month?

*iv*. Have you experienced difficulty concentrating or staying focused in the last month?

*v*. Do you feel you have more problems with memory than most people?

Before analysis of these five items, consistency in responses within person at a given visit can (should) be explored by cross-tabulating items *i* and *iv* (both assess "difficulty concentrating in the past month", with slightly different wording). If low levels of agreement across these two items is observed (where "low" should be defined as anything less than 85%, since these are essentially the same question), then the rest of the items may also be of lower reliability than would be needed to support hypothesis testing; hypothesis *generation* may still be supported by the other items but they should not be combined into a "total score". If the responses on these items agree (86-100%), then it is possible to consider the four (different) items as comprising an index of the perceived impact of cognitive symptomatology. These items cannot be combined and/or considered to directly represent the amount or extent of cognitive impact, because they are framed from the perspective of the patient's experience, and were not developed to "measure" cognitive status. Moreover, item *v* is constructed differently from the other items, so including it (even as a 0/1 no/yes item) in a simple sum of ratings would not contribute the same level/type of information as the other items would. However, recoding the other four items as 0 (frequency ratings of 0/never) or 1 (frequency ratings >0/ever) would then make all four items (i.e., because there are two instances of the same question in this set of five, items *i* and *iv*) similar in construction and the interpretability of the responses. Then a simple sum of 0s and 1s would yield a "total perceived impact score" ranging from 0-4; suggest difficulty with a variety of cognitive tasks. Higher scores could suggest a wider-ranging problem, or a problem that is perceived to have greater impact for the patient.

*3.1.2 Cognitive tasks-specific assessment*

The FI dataset also includes 15 items from a single questionnaire, which was originally designed to capture the difficulty perceived *by the caregiver* when the Parkinson's patient is doing a variety of cognitive tasks, comprising the PDAQ-15. Items on this scale are rated from 0 (no difficulty) to 4 (cannot do) (see Brennan et al. 2016), and the total (simple sum) of these items (0-60) constitutes the "total score". Importantly, in the FI data set, the items are worded to obtain the *patient's*, and not the caregiver's, point of view. Lower scores (close to zero) are best, high scores suggest increasing levels of difficulty with cognitive tasks (i.e., lower levels of cognitive functioning). The item content for all 50 items from which these 15 were taken are given in supplemental materials to Brennan et al. (2015), here: mds26339-sup-0001-suppinfo01.docx .

The 15 PDAQ-15 items ask respondents to rate "how much difficulty" they have in the following tasks:

- How much DIFFICULTY do you currently have keeping track of time (e.g. using a clock)?
- How much DIFFICULTY do you currently have counting the correct amount of money when making purchases?
- How much DIFFICULTY do you currently have reading and following complex instructions (e.g. directions for a new medication)?
- How much DIFFICULTY do you currently have handling an unfamiliar problem (e.g. getting the refrigerator fixed)?
- How much DIFFICULTY do you currently have explaining how to do something involving several steps to another person?
- How much DIFFICULTY do you currently have remembering a list of 4 or 5 errands without writing it down?
- How much DIFFICULTY do you currently have using a map to tell where to go?
- How much DIFFICULTY do you currently have remembering new information like phone numbers or simple instructions?
- How much DIFFICULTY do you currently have doing more than one thing at a time?
- How much DIFFICULTY do you currently have learning to use new gadgets or machines around the house?
- How much DIFFICULTY do you currently have understanding your personal financial affairs?
- How much DIFFICULTY do you currently have maintaining or completing a train of thought?
- How much DIFFICULTY do you currently have discussing a TV show, book, movie, or current events?
- How much DIFFICULTY do you currently have remembering what day and month it is?

These 15 items constitute everyday tasks (activities of daily living) that vary in terms of whether or not any given individual would even engage in the task (e.g., learning to use new gadgets, or counting the correct amount of money when making a purchase, which are two tasks that might not happen very often), as well in terms of how difficult the patient finds the task. Thus, it is important to check on whether items have zeros that are ratings (i.e., assessments by a caregiver that the patient has no difficulty doing the task), and are not simply missing because the individual never does that task. If analyses of cognitive status, or change in status, are contemplated, then it might be advisable to

create a subset of these tasks that have missingness less than some *a priori* threshold (e.g., no item has more than 10% missingness, meaning that at least 90% of the respondents attempt the task). The simple sum of PDAQ-15 items at a single visit will give an overall impression of the cognitive difficulty a given patient perceives themselves to be experiencing. It is important to recognize that this score is an *indirect* measure of cognitive function or status; conclusions based on this score must be contextualized as such.

Change in cognition should not be assessed using a simple change score (subtraction) because none of these items (neither the first five considered, nor those of the PDAQ-15) are exchangeable; in fact Brennan et al. (2016) specify that the 15 items in the instrument were chosen (from a pool of 50) in part because of the *range* of cognitive abilities that are included. Two of the first set are clearly exchangeable, as noted above –which is why only one of them is included in any scoring; the 19 items described here are not exchangeable. Thus, "no change" in total score may be observed when in fact, the patient has experienced an improvement in difficulty for one item –possibly due to additional practice – while worsening difficult simultaneously is experienced in another (possibly due to actual worsening of cognitive state). Changes *per item* (e.g., "has the amount of difficulty in keeping track of time increased over time?") or using a qualified change algorithm (see Tractenberg et al. 2013) should be used instead of change in total score, if changes in cognitive function are of interest.

**3.2 "OFF" episodes**

The "off" phenomenon in Parkinson's disease is defined as the situation where a patient who has symptoms that were successfully treated with medication temporarily experiences a gap in the medication's effects or effectiveness. This type of episode, also referred to as "fluctuations" (Horne, McGregor & Bergquist 2015), "wearing off" (Papapetropoulos 2012) and "hypomobility" (Obering, Chen & Swope 2006) is naturally very worrisome for patients and their caregivers, but apart from the fact that they seem to emerge over time (i.e., are not predictable from the patients' early experiences with pharmacological treatment), very little is known – or assessed - about OFF episodes. A new survey was developed and is included in the FI dataset, that "seeks to understand how patients and carepartners understand and communicate about OFF periods, and how OFF periods impact on patients and carepartners". The survey is estimated to require 25 minutes to complete, and includes 19 questions – several of which are multi-part. For investigators who want to know more about patient characteristics associated with OFF episodes and symptoms, the prevalence of this experience for Parkinson's patients, or how often patients experience these (and if that frequency increases over time), there is a *subset* of just six of these 19 items that should be considered. Each is discussed below.

*3.2.1 OFF episode items and their scoring considerations*

A. Do you experience OFF periods, as (just) defined?

This item can be scored as 0 (no OFF periods) or 1 (yes OFF periods) to generate an estimate of prevalence of OFF episodes; it can also be used as a sentinel item to identify individuals who change over time in the FI study from 0 (no) to 1 (i.e., change from not experiencing them to having experienced them).

It is essential to ensure that only responses from individuals who answered item A "yes" (1, indicating they do experience OFF episodes) are considered on the remaining items; not only because if a patient says they do *not* experience OFF episodes, but gives a non-zero/not-missing answer to any of these items *specifically about* OFF episodes, it is

unclear whether they are describing actual OFF episodes or their general symptom experience. These respondents will dilute any conclusions that would otherwise be plausible for individuals who state that they *do* experience OFF episodes (i.e., answer item A "yes").

B. Over the last week, on average, how many OFF episodes do you experience in a typical waking day?

Select one:

- No episodes, zero
- 1 episode per day
- 2 episodes per day
- 3 episodes per day
- 4 episodes per day
- Greater than 4 episodes per day
- I don't know

This item (B) can be rescored to range from zero ("no episodes") to 5 where "5" represents "greater than four episodes per day" (and any alternative number greater than 4 can also be used). "I don't know" would not receive a score (i.e., would be missing).

C. Over the last week, on average, what is the typical duration of each OFF episode?

Select one:

- Less than 15 minutes
- Between 15 and 30 minutes
- Between 30 minutes and 45 minutes
- Between 45 minutes and 1 hour
- Between 1 hour and 2 hours
- Greater than 2 hours
- I don't know

Similar to item B, this item can be rescored to range from 1 to 6 where "6" represents "greater than 2 hours" (and any alternative number greater than 6 can also be used). "I don't know" would not receive a score.

The sum of item B and C could give a general idea of burden (frequency and duration) of OFF episodes; their product could also, and would expand the range of "scores" from 0 - 11 (sum) to 0-30 (products). There are only two items, and these are rated subjectively, but responses involve very specific timings (i.e., rather than "a little, some, a lot" or other individualized ratings). It is unknown whether either the sum or product version of these "scores" would function meaningfully as continuous outcomes for clinical research or clinical trials; however, these could function as co-primary outcomes with "success" of an intervention defined as "improves by at least one level". This would render the outcomes countable (rather than use a falsely-continuous version of time), or successes could be ranked so that "two points" are awarded for "improves by at least one level" on both B and C; one "point" is awarded for "improves by at least one level" on just one; zero points are awarded for "does not improve by at least one level on either", and points are deducted if the individual worsens by one or two points.

Additional questions about these two "OFF items" (B & C) that could be explored in future work include:

- What are the ranges of sum and product "scores" on OFF symptoms?

- Are there distinctive groups of patients in the "high" and "low" score ranges?
- Do individual patients change from "low" to "high" as their Parkinson's progresses?
- Do individual patients ever change from "high" to "low" spontaneously?
- Would changes from "low" to "high" or vice versa be important or plausible as clinical or clinical trial outcomes?

Given how little is known (or published) about OFF episodes, these five research questions, targeting just the two patient-reported outcome items B and C, could be helpful in understanding patient experiences of these episodes, as well as informing clinical trial design to prevent or mitigate these episodes. All of these research questions can be addressed with the FI data once it becomes available.

D. One item on this survey that asks respondents to indicate whether or not (yes/no/not sure) their OFF episode includes any of a list of 24 different symptoms. For an individual who experiences multiple episodes with *different symptoms* "emerging" or "breaking through" their medication, interpreting this item could be complicated. The survey does not include items that ask whether the OFF episode symptom experience is actually a symptom that they did/do have that is actually controlled (usually) by medication, or if instead they experience different (additional) symptoms during these episodes or just the ones that they're being treated for. However, responses on this item could be useful to explore whether there are a set of symptoms, whether they breakthrough (occur when typically they are controlled) or emerge (only occur in OFF episodes but never otherwise), that are typical of the OFF phenomenon. Technically this item could be scored from 0-24, but that would not be representative of the same experience for all respondents – so that scores of equal "number" would not actually represent the same level of breakthrough symptomatology. Also, such scores could be low either because a patient is reporting their first ever experience or because they have had the experience repeatedly but it is being controlled well/better. This item might be better (or, useful only) as an indication of change – e.g., to indicate if additional, or different, symptoms are experienced over time. Each patient would then be scored a countable, and exchangeable "yes" if responses on subsequent visits have a greater number of the list of 24 symptoms endorsed, or if different symptoms are endorsed (but the same number as a prior visit). This item will be difficult to combine with other OFF episode items into a single score, but profiles of symptoms or of changes (e.g., Tractenberg et al. 2007; Tractenberg et al. in preparation) could be used to described and classify patients, and used as a categorical outcome.

E. There are two additional items that are of potential interest and importance to developing a better understanding of OFF episodes and their impact as perceived by the patient: asked as separate items, patients "rate their disability" during ON (when the medication is working as intended) and during OFF (temporary disruption in the medication working as intended) states. Each individual's *relative* ratings should be considered together, due to well-known biases that self-report can create (see, e.g., Tractenberg et al. 2013). Thus, a *ratio* of each individual's disability ratings should be used. The arrangement of ratings (i.e., ON/OFF or OFF/ON) would depend on the investigator's interest, and would also support hypothesis *generation* rather than testing, because the items were not designed specifically to provide an index of the perceived impact of OFF episodes relative to the individual's "normal" (i.e., ON) perception of their disability. If the ratio is formed as "ON/OFF", the interpretation would be along the

lines of, “how much disability is experienced when the medication is successful (compared to OFF levels, not the general public or disability prior to diagnosis)?”. If the ratio is formed as “OFF/ON”, the interpretation would be “how much more burden (than when symptoms and disability are controlled) does an OFF episode create?”. These interpretations could be studied by estimating levels of symptomatology in newly-diagnosed individuals or those experiencing OFF episodes for the first time. They could also be used to study the range of individual experiences of how the OFF episodes affect patients at different severity levels or along the natural history of the disease. Because of the very divergent meanings of the two ratios, it is essential to explicitly describe the analysis of these two items in the methods, and to fully report all analyses that are done to study the relationships between the patients’ perceived burden or disability and any other variables or scores relating to either OFF episodes (e.g., item D) or to other features of Parkinson’s disease.

Finally, once scoring (for any or all of this subset of the OFF episode survey items) is validated, a version of the instrument that features some combination of these items (B-E) could be highly useful in clinical trials and other research for which the range of outcomes is currently dominated by diaries (which cannot be used in research earlier in the translational research continuum), and duration of the breakthrough symptoms (which can be used in research earlier in the translational research continuum, but may not be useful in understanding the causes and mechanisms of the OFF episode). It is unlikely that a single value can effectively summarize the wide range of information that these OFF episode survey items comprise, but profiles or patterns can be developed to characterize (rather than quantify) patient experiences with this phenomenon. More research (in addition to formal psychometric analyses, discussed in the next section), is required before the OFF episode items can be used reliably and validly in clinical research.

### 3.3 COSMIN criteria

Table 1 presents the alignment of the COSMIN criteria and the two types of outcomes (cognitive status; OFF episodes) described here.

**Table 1. COSMIN criteria and the Fox Insight (FI) dataset outcomes based on cognitive status and OFF episode items**

| *COSMIN construct:* | *Definition (see Tractenberg et al. 2018)* | *For FI-based assessment of Cognitive status* | *For FI-based assessment/study of OFF episodes* |
|---|---|---|---|
| **Reliability-internal consistency** | Degree of interrelatedness among items | PDAQ-15 has published reliability data (Brennan et al. 2016). | NA – items cannot be combined in “scores” |
| **Validity-content** | Degree to which instrument measures the construct it targets | PDAQ-15 has published content data (Brennan et al. 2016). | Research is required to determine |

| | | | |
|---|---|---|---|
| **Validity – face** | Degree to which items "look" as if they are an adequate reflection of the target construct | By development & design, iteratively eliciting and obtaining input from patients and clinicians. | Research is required to determine-but the FI dataset may provide some information. |
| **Validity-construct** | Degree to which the scores are consistent with expected similarities (convergent) and differences (divergent) between groups | PDAQ-15 has published construct data (Brennan et al. 2016). | NA – items cannot be combined in "scores" |
| **Validity-criterion** | Degree to which the scores reflect a "gold standard" | PDAQ (50 items) has published construct data (Brennan et al. 2015). | Research is required to determine-but the FI dataset may provide some information (particularly with item A). |
| **Validity-structural** | Degree to which the scores are an adequate reflection of the dimensionality of the target construct | PDAQ (Brennan et al. 2015) and PDAQ-15 (Brennan et al. 2016) have published construct data | NA – items cannot be combined in "scores" |
| **Interpretability** | Degree to which a qualitative meaning (patient/clinician perspectives) can be given to the scores | Detailed descriptive statistics | Detailed descriptive statistics |

The new outcomes we discussed, based on the FI dataset, do not support estimation of internal consistency (COSMIN defined "reliability"), either because they have already been estimated elsewhere (e.g., for the PDAQ-15 by Brennan et al. 2016) or because this feature is not applicable to the suggested scoring of these outcomes (OFF items). Similarly, COSMIN defined "construct validity" as representing the dimensionality with which an instrument captures the construct of interest. Since the OFF episode items do not lend themselves to "scores", this COSMIN criterion could not be met in the same way as a psychometrically-derived instrument would (see Tractenberg et al. 2018). Similarly, structural validity is also not estimable for any version of the OFF episode items or scores discussed here. Similarly, "responsiveness" is defined by the COSMIN report as reflecting the sensitivity of an instrument to change in the construct under study. Due to the difficulties in defining "cognitive changes" associated with Parkinson's disease, and the range of what is not yet known about OFF episodes, even defining "the construct under study" is highly challenging. The FI data are longitudinal, but no clinical or objective endpoints, or variables that could be considered milestones, are currently included. Thus, no gold standard for clinically meaningful change on any patient experience or other dimension of Parkinson's disease exists in the dataset (although there are some initiatives to link *at least some* of the FI data to genetic and other biomarker data). In fact, creating reliable and valid representations of clinically meaningful states,

and changes in those states, is our objective in exploring and describing these two new outcome types (cognitive function/OFF episodes); these will need to be further validated in the future with independent samples where clinical outcomes *are* included. Thus, the COSMIN criteria *can* be evaluated, based on scores that originate with the FI data – but not with the FI data itself.

## 4. Discussion

The international *Co*nsensus-based *S*tandards for the selection of health *M*easurement *In*struments (COSMIN, Mokkink et al. 2010) gives four essential psychometric features of health related patient reported outcome (PRO) by international agreement: *Reliability* (minimization of measurement error; internal consistency or interrelatedness of the items; and maximization of variability that is due to "true" difference between levels of the symptoms across patients); *validity* (*content*, reflection of the construct to be measured; *face*, recognizability of the contents as representing the construct to be measured; *structural*, the extent to which the instrument captures recognizable dimensions of the construct to be measured; and *criterion*, association with a gold standard); *cross-cultural validity*; and the *interpretability of scoring*.

It has been argued (Tractenberg et al. 2018) that instruments should have established item level face, content, and possibly cross-cultural validity before structural validity and reliability can be estimated. For the OFF episode "outcomes", several scientific questions remain to be answered before construct or structural validity can be studied. The interpretability of the scoring suggestions here are also hypotheses that can be tested empirically, although not with the FI data in all cases. For the cognitive status "outcomes", while a total score has been empirically demonstrated to have many of the COSMIN psychometric criteria (and indeed, the 15 items (Brennan et al. 2016) were selected from the larger pool (Brennan et al. 2015) based on item level psychometric characteristics), interpretability of *change in total scores* is unlikely to support the use of total scores in clinical trials (because estimating change would need to take the items into account; Tractenberg et al. 2013). Specifically, these items are not exchangeable. A variety of empirically testable questions arise from the scoring considerations for both cognitive and OFF items. In addition to enriching the value and utility of the FI dataset to the research community, addressing the psychometric considerations of these new outcomes as outlined in the Results section could facilitate clinical research into Parkinson's disease.

Patient reported epidemiological data are becoming more widely available (e.g., https://www.samhsa.gov/capt/practicing-effective-prevention/epidemiology-prevention/finding-data ; https://researchguides.uic.edu/c.php?g=252253&p=1683071 ; see Packer, 2016); so while these results are useful for planning analyses of the FI dataset, and can also support new empirical tests of these hypotheses to enrich our understanding and ultimately, the search for treatments for Parkinson's disease, these new sources and scoring recommendations may also may support future designs for similar patient reported epidemiological data sets, so that they include outcome variables or variables that can be used to demonstrate alignment of derived outcomes with the COSMIN validity criteria. However, it is essential that investigators and analysts understand the limitations that more informally-collected data can pose for hypothesis testing and the reproducibility of results that arise from analyses that do not incorporate clinical gold standards. While more research is needed to better understand OFF episodes, because there are no clinical gold standards available as yet (2018), much more

is known about cognitive function in Parkinson's and other movement disorders (see, e.g., Marras et al. 2014; Robbins & Cools, 2014). While there are plans to link available genetic and biomarker data on Parkinson's patients who have contributed (or will contribute) to the FI dataset and also to databases on biomarkers and genetics (e.g., 23andMe), without formal cognitive assessments, results based on the FI cognitive items should be considered more hypothesis *generating* than hypothesis *testing*. It is essential – particularly for machine learning and other algorithmic approaches to the FI data – that the scoring limitations and non-assessability of the COSMIN criteria - be recognized and kept in mind as results are interpreted. Sensitivity analyses will be critical to support any conclusions about cognitive status, change, or OFF episodes derived from the items discussed here. Independent replications would be needed for conclusions about cognition in Parkinson's that can support clinical trials or theory building/testing.

## Acknowledgements

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.

## References

Antrobus E, Elffers H, White G, Mazerolle L. Nonresponse bias in randomized controlled experiments in criminology: Putting the Queensland Community Engagement Trial (QCET) under a microscope. *Eval Rev*. 2013 Jun-Aug;37(3-4):197-212. doi: 10.1177/0193841X13518534.

Bollen, K. A. *Structural equations with latent variables*. (Wiley, 1989).

Brennan L, Siderowf A, Rubright JD, Rick J, Dahodwala N, Duda JE, Hurtig H, Stern M, Xie SX, Rennert L, Karlawish J, Shea JA, Trojanowski JQ, Weintraub D. Development and initial testing of the Penn Parkinson's Daily Activities Questionnaire. *Mov Disord.* 2016 Jan;31(1):126-34. doi: 10.1002/mds.26339.

Brennan L, Siderowf A, Rubright JD, Rick J, Dahodwala N, Duda JE, Hurtig H, Stern M, Xie SX, Rennert L, Karlawish J, Shea JA, Trojanowski JQ, Weintraub D. The Penn Parkinson's Daily Activities Questionnaire-15: Psychometric properties of a brief assessment of cognitive instrumental activities of daily living in Parkinson's disease. *Parkinsonism Relat Disord*. 2016 Apr;25:21-6. doi: 10.1016/j.parkreldis.2016.02.020.

Egleston BL, Miller SM, Meropol NJ. (2011). The impact of misclassification due to survey response fatigue on estimation and identifiability of treatment effects. Stat Med. 2011 Dec 30;30(30):3560-72. doi: 10.1002/sim.4377.

Fielding S, Fayers P, Ramsay CR. Analysing randomised controlled trials with missing data: choice of approach affects conclusions. Contemp Clin Trials. 2012 May;33(3):461-9. doi: 10.1016/j.cct.2011.12.002.

Halbesleben JR, Whitman MV. Evaluating survey quality in health services research: a decision framework for assessing nonresponse bias. Health Serv Res. 2013 Jun;48(3):913-30. doi: 10.1111/1475-6773.12002.

Hansen RA, Henley AC, Brouwer ES, Oraefo AN, Roth MT. Geographic Information System mapping as a tool to assess nonresponse bias in survey research. Res Social Adm Pharm. 2007 Sep;3(3):249-64.

Heavner K, Newschaffer C, Hertz-Picciotto I, Bennett D, Burstyn I. (2014). Quantifying the potential impact of measurement error in an investigation of autism spectrum disorder

(ASD). J Epidemiol Community Health. 2014 May;68(5):438-45. doi: 10.1136/jech-2013-202982.

Horne MK, McGregor S, Bergquist F. (2015). An objective fluctuation score for Parkinson's disease. PLoS One. 2015 Apr 30;10(4):e0124522. doi: 10.1371/journal.pone.0124522. eCollection 2015.

Mokkink, L. B. *et al.* The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J. Clin. Epidemiol.* **63,** 737–745 (2010).

Molinari NM, Wolter KM, Skalland B, Montgomery R, Khare M, Smith PJ, Barron ML, Copeland K, Santos K, Singleton JA. Quantifying bias in a health survey: modeling total survey error in the national immunization survey. Stat Med. 2011 Feb 28;30(5):505-14. doi: 10.1002/sim.3911

Nunnally, J. C. & Bernstein, I. H. *Psychometric theory*. (McGraw-Hill, 2008).

Obering CD, Chen JJ, Swope DM. (2006). Update on apomorphine for the rapid treatment of hypomobility ("off") episodes in Parkinson's disease. *Pharmacotherapy*. Jun;26(6):840-52.

Oleson JJ, He CZ. Adjusting nonresponse bias at subdomain levels using multiple response phases. Biom J. 2008 Feb;50(1):58-70. PubMed PMID: 17849386.

Packer M. (2016). Data sharing: lessons from Copernicus and Kepler. *BMJ* 2016; 354 doi: https://doi.org/10.1136/bmj.i4911

Papapetropoulos SS. (2012). Patient diaries as a clinical endpoint in Parkinson's disease clinical trials. CNS Neurosci Ther. 18(5):380-7. doi: 10.1111/j.1755-5949.2011.00253.x.

Pierce BL, VanderWeele TJ. The effect of non-differential measurement error on bias, precision and power in Mendelian randomization studies. Int J Epidemiol. 2012 Oct;41(5):1383-93. doi: 10.1093/ije/dys141. Erratum in: Int J Epidemiol. 2014 Dec;43(6):1999.

Tavakol, M. & Dennick, R. Making sense of Cronbach's alpha. *Int. J. Med. Educ.* **2,** 53–55 (2011).

Tractenberg RE, Groah SL, Rounds AK, Ljungberg IH, Schladen MM. (2018). Preliminary validation of a Urinary Symptom Questionnaire for individuals with Neuropathic Bladder using Intermittent Catheterization (USQNB-IC): A patient-centered patient reported outcome. *PLoS One*. 2018 Jul 10;13(7):e0197568. doi: 10.1371/journal.pone.0197568. eCollection 2018.

Tractenberg RE, Groah SL, Rounds AK, Davis E, Ljungberg I, Schladen M. (in preparation). Scoring and validity in patient-centered PROs: A case analysis with the urinary symptom questionnaires for neurogenic bladder.

Tractenberg RE, Weiner, M. F., and Chuang, Y-L. (2007). CES-D symptoms and DSM criteria: Levels of depressive symptomatology in a cohort of Taiwanese elderly. *Journal of Chinese Clinical Medicine* 2(1): 1-10.

Tractenberg RE, Yumoto F, Aisen PS. (2013). Detecting When "Quality of Life" Has Been "Enhanced": Estimating Change in Quality of Life Ratings. *Open J Philos*. 2013 Nov 1;3(4A):24-31.

Zheng HW, Brumback BA, Lu X, Bouldin ED, Cannell MB, Andresen EM. Doubly robust testing and estimation of model-adjusted effect-measure modification with complex survey data. *Stat Med*. 2013 Feb 20;32(4):673-84. doi: 10.1002/sim.5532