Machine Learning to Evaluate the Quality of Patient Reported Epidemiological Data

Futoshi Yumoto[1,2,5], Robert L. Wood[1,3], Rochelle E. Tractenberg[2,4,5]
[1] Resonate, Inc., Reston, VA
[2] Collaborative for Research on Outcomes and –Metrics; Silver Spring, MD
[3] Wichita State University, Wichita, KS
[4] Departments of Neurology; Biostatistics, Bioinformatics & Biomathematics; and Rehabilitation Medicine; Georgetown University, Washington, D.C.
[5] Psychometrics Core, Fox Insight Study (FI); The Michael J. Fox Foundation for Parkinson's Research

Abstract
Patient reported epidemiological data are becoming more widely available. One new such dataset, the Fox Insight (FI) project, was launched in 2017 to encourage the study of Parkinson's disease and will be released for public access in 2019. Early analyses of responses from the earliest participants suggest that there may be significant fatigue effects on elements that occur later in the surveys. These trends point to potential violations of assumptions of missingness at random (MAR) and completely at random (MCAR), which can limit the inferences that might otherwise be drawn from analyses of these data. Here we discuss a machine learning approach that can be used to evaluate the likelihood that an individual respondent is "doing their best" vs. not. Bayesian network structural learning is used to identify the network structure, and data quality scores (DQS) were estimated and analyzed within- across-each section of a set of seven patient reported instruments. The proportion of respondents whose DQS scores fell below what would be considered a cutoff (threshold) for data that is unacceptably or unexpectedly similar to random responses ranges from a low of 13% to a high of 66%. Our results suggest that the method is not unduly influenced by the length of instruments or their internal consistency scores. The method can be used to detect, quantify, and then plan or choose the method of addressing nonresponse bias, if it exists, in any dataset an investigator may choose – including the FI dataset, once that is made available. The method can also be used to diagnose challenges that may arise in one's own dataset, possibly arising from a misalignment of patient and investigator perspectives on the relevance or resonance of the data being collected.

Key Words: Machine Learning, data quality, Bayesian Network, mutual information, trustworthiness of data, data assessment.

1. Introduction

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). 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).

1.1. Fox Insight (FI) dataset

One new such dataset, the Fox Insight (FI) project, was launched 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/). To date (May 2018), 12,000 individuals with Parkinson's disease and 4,000 controls have contributed data to this resource. The resource will be made public in July 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). In the FI study, 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. However, study "visits" (i.e., opportunities to complete the online questionnaires) can require from 15 to 60 minutes, and preliminary analyses of responses from these earliest participants in the FI study suggest that there may be significant fatigue effects, such that individuals are more likely to complete more of the questions/instruments that are administered earlier in a "visit" than they are to answer questions that appear later in the visit. Because a great deal of personal and health related data are collected for each participant and at each visit, the data are anticipated to be of high quality. Not all resources that researchers want to utilize will be of such high quality, and when investigators merge data from across different sources, the quality of the resulting dataset will be difficult to estimate.

1.2 Fatigue and data quality in longitudinal data on Parkinson's disease

Fatigue is a significant clinical symptom in PD, and "response fatigue" is also a significant problem in any survey-based research (Egleston et al. 2011). These two types of fatigue, both being prevalent in the FI data set, may result in potential violations of assumptions of missingness at random (MAR) and completely at random (MCAR), which can limit the inferences that might otherwise be drawn from analyses of these data. Practically, they also affect outcomes and decisions that are based on analyses of such data (Egleston et al. 2011; Fielding et al. 2012; Zheng et al. 2013; see also Pierce & VanderWeele, 2012; Heavner et al. 2014). Although many methods have been described recently to address the effects of the bias such missingness can create in inferences and conclusions based on survey data (for example, Hansen et al. 2007; Oleson & He, 2008; Molinari et al. 2011; Antrobus et al. 2013; Halbesleben & Whitman 2013), none have addressed how investigators who seek to take advantage of the ever-increasing availability of large-scale epidemiologic data such as that in the FI study to determine whether either clinical fatigue, response fatigue, or both are present in the dataset before beginning data analyses. In a more general sense, before investing additional time and resources in analyzing a large or massive set of data, the quality of that data should be evaluated- irrespective of the method that would be used to accommodate or overcome difficulties arising from missingness mechanisms. This paper discusses a novel method for doing this. Specifically, the method can be used to evaluate the likelihood that an individual respondent is "doing their best" vs. not on any given visit (where "visit" is defined as the collection of responses at one timepoint). The method is described and demonstrated using the current (May 2018) data in the FI study database.

1.3 Methodological considerations to identify fatigue and other limiters of data quality

The method to evaluate data quality generates an estimate of data quality from each respondent; the estimate can describe (or summarize) each section of the dataset, e.g., by instrument; or over all data, e.g., for a respondent's entire contribution. The estimate can be used to identify and estimate fatigue effects that can arise when a survey is too long or

where participants' contributions may change from "informative" (high quality) to "less informative" (lower quality than is typical for a respondent or for the group on average).

The method leverages relationships in the dataset itself, which can be modeled using mutual information. The mutual information (MI) of two discrete random variables, X and Y, in a given dataset, can formally be defined formally as follows:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

where p(x) and p(y) is the marginal probability distributions of X and Y, and p(x,y) as the joint probability distribution of X and Y. Informally, mutual information quantifies the question, "given that we know the state of Y, how much does that tell us about the state of X?" In the context of a survey, if you ask a respondent what is their favorite color (example question X), the response will provide little to no information about the market value of their house (example question Y). By contrast, if you ask what neighborhood they live in (example question X'), the response could likely entail information about the potential market value of their house (example question Y) – for example, knowing the answer to X' could force the answer to Y into a fairly narrow range. While 'favorite color' and 'market value of house' have little to no mutual information, 'neighborhood' and 'market value of house' have higher/high mutual information. This method can be used to assess response quality from variables from multiple, unrelated surveys already within one dataset, as well as for the results of large-scale data capture, or a combination of these. Examples may arise in business, biomedical or healthcare datasets; or epidemiologic studies. Across these contexts, the method will work whether the mutuality in the information comes from one person giving responses to items (even if these items are from unrelated surveys) or, whether people &/or their contexts give rise to some level of mutuality in the data.

1.3.1 Badness in data and data quality

Any given dataset presumably has some "bad" data – where "bad"ness arises from the respondent not doing their best, being tired, inattentive, confused, etc., or, based on errors in the non-survey data, including out of date records and other errors for some of the observations in the dataset. "Bad" data, so defined, is assumed to be present in any dataset, albeit at an unknown level. The "bad data" exerts diluting influences on estimations that would be computed based on the data, making estimates less precise, or biasing the estimate (as well as affecting the precision). Although any data will have "badness" to some unknown extent, the dataset will still have a detectable level of mutuality of the information. If all respondents are contributing "bad" data (and/or if all the values are wrong), the data would be completely random (and MI = 0); if all respondents are perfect/all observations correct, then whatever level of mutual information that is predicted for each instrument, survey, or data set should be observed/recovered. Since data are never perfect, the best (highest) value the MI could plausibly achieve in any case would effectively be the lower bound on the "true" mutual information. That is, the true level of MI, based on the variables (rather than the observations) in the dataset, could be higher than the MI that is actually observed for the given dataset; if the respondents and their data were perfect, the MI based on a perfect version of the data would not be lower than the MI derived from what is observed. Thus, for any given dataset, MI can be estimated, but it will not be clear whether this is value will be lower than what it should be ("true" MI) simply because data are never perfect *or* if it is lower than what it should be because of some unknown level of "bad" data.

1.3.2 Mutual information and data quality

The challenge, then, is to determine what level of MI is needed in order to support a conclusion that the data are *trustworthy*. This method works to identify the upper bound of mutual information that would be obtained, given your data structure, if the data were "bad", but not completely random. That is, the method essentially estimates the *highest* MI you can have for bad data (assuming that the lowest MI is effectively zero). The resulting data quality score, described in detail below, therefore suggests whether you can trust the data: conclusions based on the data would only be supported if the data quality score suggests that the minimum acceptable MI value has been met or exceeded.

In cases where the average mutual information between items is below the MI threshold estimated by this method, background knowledge, simulation studies, and other sensitivity modeling might be required to build better understanding of which variables are or should be observed to have high/higher mutual information. The MI threshold will depend critically on both the context of the data that was collected (e.g., marketing or analytics context, vs a biomedical or epidemiological context) and the known psychometric properties of instruments being used to collect the data. For example, the MI observed for a massive dataset assembled from multiple sources (e.g., by scraping) is by its nature going to be lower than the MI in a large or massive dataset that is designed for a clinical, biomedical, or epidemiological application. The upper bound on unacceptable MI (i.e., the MI threshold) is identified by comparing the observed data to a fully- random, but otherwise parallel, version of your dataset (see method section for more detail).

This method utilizes machine learning to find all relevant mutual information relationships in a given dataset, where "relevant" is defined as relationships between any pair of variables where $I(X;Y)>0.05$ (i.e., where at least 5% of information is mutual). This provides a strong probabilistic position for assessing the observations pertaining to a survey taker or respondent in that dataset, but can also be used to assess the data itself. If the joint probability of an individual's responses throughout the dataset is relatively low as compared to other respondents, then that respondent is either idiosyncratic (i.e., a legitimate statistical outlier), or may be demonstrating fatigue (or inattention) beyond what is expected (or what can be tolerated in terms of the bias that they may impart to any estimates). Conversely, if there is an overall low level of MI, it may suggest weak or non-representative sections of the dataset (e.g., incorrect values, poorly-functioning surveys, etc.).

2. Overview of the data quality score (DQS) estimation method

The data quality can be estimated through a system whereby two scores are created, one representing the lowest acceptable quality, derived from *random* (simulated) responses, and one representing "actual" quality, derived from *observed* responses. The system is generally described by these steps:

1. Unsupervised structural learning uses MI and known features of subsets of questions (e.g., instrument characteristics including # of questions, categories of responses, etc., if these exist) to quantify the associations among all variables in each section (or the whole dataset, as appropriate) and derive the underlying network structure.
2. Using the same known features of subsets of questions used to derive the network structure in step 1 from the observed data, generate random responses that reflect the specifications of the survey or instrument (or use the bounds of

the variables, e.g., income, location, etc. if the data are scraped from non-survey source). Random responses do not utilize the network, only the subsetting/instrument features.

3. The unconditional probabilities, representing the **denominator** of the individual's *observed* Data Quality Score (DQSo), are calculated as a function of each individual's observed responses without any information about the network, surveys/instruments, or variables).
4. The likelihood of each participant's responses within each section/instrument (or the whole dataset) is calculated as a function of the conditional probabilities given the network structure derived in Step 1, which includes consideration (and leveraging the mutual information of) all X/Y pairs found in Step 1, as well as specifications of the survey/instrument/variables. The difference between the individual's conditional and unconditional probabilities represents the **numerator** of their *observed* Data Quality Score (DQSo).
5. Take the random data from step 2 to calculate unconditional probabilities (Step 3) but for random data.
6. Take the random data from step 2 and calculate the conditional probabilities of the random data given the network structure from Step 1 (fit should be very poor) (Step 4). Thus, steps 3 and 4 are repeated using the random data, and are used to create a *random* Data Quality Score (DQSr). Both distributions of DQSs will tend to follow a normal (bell-shaped) curve. The distributions of DSQs from observed and random data can be superimposed. The DSQos that overlap with DSQrs will be the "diluting" responses – capturing survey takers who have a significant number of responses that appear to be unrelated – apparently-random, to an unexpected extent. These respondents are characterized, based on their DSQos as having observations in the dataset that appear more independent and disconnected to other answers they have already given than is observed for the majority of others in the dataset. The DSQos at the high end of the distribution reflect "expected" response types, responses that are very likely, common, and/or very different from random.

2.1 Using/interpreting the DQS

The DQS scores for random and for observed data can be used for different purposes. The 95^{th} percentile of the DQSr distribution (i.e., the score below which 95% of all random data DQSr fall) can be taken as a threshold, below which no DQSo (i.e., observed) will be considered acceptable. This would then define "bad data" – any respondent with a DQSo below the DQSr 95^{th} percentile would be considered "random". To generate an overall estimate of how much "bad" data is included in a given dataset, the proportion of respondents whose DQSo fall below the DQSr threshold would be defined as that proportion. Further, individuals whose DQSo fall below the DQSr threshold would be potentially identifiable as "tired", "inattentive", or otherwise, not doing their best. Finally, examining the DQSo for successive segments of a long/multi-part survey set could help identify places in the survey where respondents tend to do worse, lose attention, "try less hard", or become fatigued.

If the cluster of DQSo that fall below the DQSr threshold (the 95^{th} percentile) represents 50% of the sample or more, it may indicate fundamentally uninformative (unexpectedly unrelated) features of the dataset. For investigators who seek public data for their research, a minimum level of quality could be specified *a priori*, to help ensure meaningful (rigorous, reproducible) estimates and conclusions that are based on whatever data are ultimately analyzed.

The system described here can be used to identify (e.g., for subsetting or sensitivity analyses) or eliminate (e.g., for "complete case analyses" that exclude individuals for whom the data are deemed "unacceptably random") the "bad" responses in/from a data set. Conversely, this system can assign a score to each individual or their responses (overall, or on sections of the dataset) according to their likelihood of being less-than-best effort data. Data quality can also be scored for each section of the dataset, collapsing across respondents (e.g., by instrument). Either of these new scores can be used to assess data quality over the duration of an individual's participation, e.g. to identify and estimate fatigue effects, as functions of the participants or as functions of sections of the dataset that may cause fatigue or inattention. In the case where fatigue effects are of interest, these can be tracked longitudinally as data permit so that fatigue effects' worsening over time can be detected.

2.2. Deriving the DQS

Step one of the method requires the creation of a network. Bayesian networks are a specific subset of graphical models. Graphical models as defined in Pearl & Russel (2000) help clarify the relationships among variables with conditional dependencies. In the graph, nodes (N) represent variables, and direct conditional relationships between variables are represented by an edge (E) between their nodes. For example, Figure 1 shows a model (M) with three variables or nodes, A, B, and C. The edges describe the relationships between them. The direction of the edges indicate causal flow of information (see Pearl & Russel, 2000).

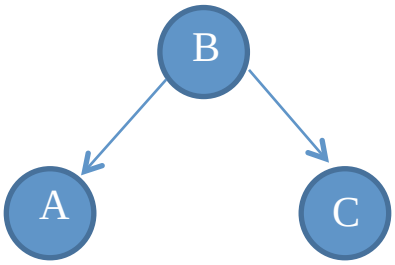

Figure 1: Example of conditionally dependent nodes in a simple graphical model.

The model in Figure 1 shows a model (M) with three nodes, or variables, A, B, and C. This simple model shows that A and B are conditionally dependent, and C and B are conditionally dependent, but as there is no connection between A and C, the variables represented by those nodes are conditionally *in*dependent, given B. In addition to these relationships among pairs of variables, it can also be said that B is a parent of A and C (Neapolitan, 2003). To be clear, it may not be the case that A and C are actually independent of one another; this figure shows that every/any way that C is impacting A is encoded in the state of B. There may be error, or some variability in A and in C that is not fully explained (or caused) by variations in B; only the conditional independence and dependence associations among and between the variables are shown in Figure 1.

The graphical model M in Figure 1 can be defined mathematically as M = {N,E,P} where N is the set of nodes, each of which represents a random variable under consideration, E is the set of edges between the nodes representing the direct conditional relationships involved, and P is the set of probability distribution functions associated with each node or variable. The same notation would apply for larger systems of variables. In Figure 1, M is the model shown with nodes (N)=A, B, C and the edges B->A and B->C. No

probability distribution is shown in Figure 1. An important aspect of this work is that the probabilistic distributions that define the relationships in the model are discoverable using software that can infer the conditional dependencies and independencies amongst a set of variables; a popular method of detection is the "method of tetrads". Given a set of four observed variables (x_1, …, x_4) and a single latent variable ξ_1, and assuming that the average disturbance is zero (i.e., $E(\delta_i)=0$ for all four indicators and these disturbance terms are all mutually orthogonal (i.e., $COV(\delta_i, \delta_j)=0$ for any $i \neq j$) and are orthogonal to the latent variable (i.e., $COV(\xi_1, \delta_j)=0$ for all i), then the covariances among the observed variables in the population (σ_{ij}) would be computed as $\sigma_{ij} = \lambda_{i1}\lambda_{j1}\phi$, where ϕ is the variance of the latent variable ξ_1 (this assumes that the factor loadings were available). If the model generating these factor loadings and factor variance is correct, then the following equations must hold in the population:

$$\sigma_{12} \times \sigma_{34} - \sigma_{13} \times \sigma_{24} = 0$$

$$\sigma_{13} \times \sigma_{42} - \sigma_{14} \times \sigma_{32} = 0$$

$$\sigma_{14} \times \sigma_{23} - \sigma_{12} \times \sigma_{43} = 0$$

Spearman (1904; 1927) discovered these relationships; Kelley (1928) labeled these tetrad differences, or tetrads, $\tau_{ghij} = \sigma_{gh}\sigma_{ij} - \sigma_{gi}\sigma_{hj}$. When $\tau_{ghij} = 0$, the tetrad is called "vanishing" or is said to have vanished, and a vanishing tetrad means there is a common cause (of the observed covariances). Tetrads are computed as the determinants of all 2x2 covariance (sub)matrices (i.e., there must be at least four variables; in variance-covariance matrices larger than 2x2, determinants are computed for every set of four variables in the matrix) (Glymour et al. 1987; Bollen 1990; Bollen & Ting, 1993). Any system of four observed variables –if they are associated (non-zero correlations)-will imply a set of three tetrad equations. These relationships are encoded with either correlations or mutual information; and thus can exploited in software that can "learn" an underlying network, which is what we used. In systems of more than four variables, all sets of four are sequentially analyzed to derive all possible tetrads.

3. Methods

3.1 Data

This study analyzed a portion of the dataset currently being collected through the Fox Insight (FI) project of the Michael J. Fox Foundation (https://foxinsight.michaeljfox.org/). These data are slated for release in 2019. Of the 24 surveys/instruments, there are 8 patient reported outcome (PRO) surveys where all respondents answer the same questions, shown in the table below. Responses from the first visit of the participant were analyzed. Not fewer than 6,000 observations on each PRO were obtained.

As described above, this study analyzed the dataset collected by Fox Insight (FI) project (https://foxinsight.michaeljfox.org/). A simulated dataset was created based on the FI survey structure, but based on the assumption that all items are independent (i.e., $I(X;Y) = 0$) to establish a baseline/"worst case" data quality score representing completely random responses.

There are 24 separate "survey" parts to an individual patient's data file in the FI dataset. Healthy controls complete 20 instruments and PD patients complete an additional 4. Of these, there are surveys about the individual's and their families' health histories and

medications, one survey that asks the individual to list (in their individualized order, and in free response) the symptom that is "most bothersome" (then the next most bothersome, etc.), and surveys that are only completed once (e.g., handedness, environmental exposure history). Of the 24 surveys/instruments, there are seven patient reported outcome (PRO) surveys where all respondents answer the same questions. Table 1 shows the PROs our analyses were focused on.

Table 1: Seven PRO instruments from the FI dataset that were analyzed

Instrument	***Topic/target of the instrument***
Parkinson's Disease Questionnaire (PDQ-8)	Functioning and well being
Euroquol Health-Related quality of life (EQ-5D)	Generic health status
Physical Activity Scale for the Elderly (PASE)	Physical activity inventory
MDS Unified Parkinson's Disease Rating Scale (MDS-UPDRS)	Movement symptoms ratings
Impact of OFF Episodes	OFF episode impact on quality of life
Parkinson's Daily Activities Questionnaire-15 (PDAQ-15)	Dependence performing cognitive tasks
Geriatric Depression Scale (GDS)	Depression/mood

Responses from the first visit of the patient participants were analyzed (i.e., we did not include the responses from healthy controls). Not fewer than 6,000 observations on each PRO were obtained.

3.2 DQS estimation

This method applies Bayesian Network methods to estimate mutual information relationships. All analyses were conducted with BayesiaLab v 7.1 (Conrady & Jouffe, 2015).

We used the Maximum Weight Spanning Tree (MWST) structural learning algorithm, due to the simplicity of the resulting learned model (MWST only allows one parent node, simplifying interpretability) for the development/discovery of a Bayesian Network that captures the expected mutual information relationships across all items in the survey, and/or per section or instrument in the survey. The network will have the features of the model shown in Figure 1, but will be much more complicated as the number of variables grows –according to the structure which is learned by the algorithm. An interim validation step confirms the structure of the network that is learned, by perturbing the data and re-learning the structure. Then, once the network structure has been confirmed in the validation step, the probability of each respondent's set of responses to all items on the survey, and separately per instrument/section of the survey, is computed using the equation below as it applies to the learned structure. This probability is the Data Quality Score, DQS, for the respondent X over the set of items (for a section or instrument in a survey, or for a full set of items, the entire survey) ranging from 1 to *n*, where *n* represents the number of items over which the probabilities are estimated.[1] By scaling the

[1] The result of this equation could serve as a slightly less precise estimate/characterization of the survey data quality. The precision can be improved if the results are scaled by the probability of the response set (Xi from 1 to *n*) assuming all questions are independent.

formula (below) by the probability of totally independent responses, each individual's DQS is not automatically decremented simply because the individual has provided less likely (than average) answers to some questions in the range from 1 to *n*.

$$DQS(X) = \frac{\prod_{i=1}^{n} P\left(X_i \middle| parents(X_i)\right) - \prod_{i=1}^{n} P(X_i)}{\prod_{i=1}^{n} P(X_i)}$$

Random data was generated by taking each of these instruments using the survey design specifications as described above and published for each instrument (if available) or detailed within the FI data dictionary. The random data include all non-open ended questions on each instrument. The data generation did not need to account for any rules on the survey (e.g. skip logic), but if this method were to be applied to data involving surveys with distinct skip patterns, those would need to be modeled carefully so as to avoid impossible logical combinations of items in estimating data quality.

3.3 Analyses of DSQ results

In order to demonstrate the utility of DQS for evaluating the quality of the set of responses that an individual provides on any given survey, we derived the DQSo from a set of patient reported outcomes (PROs) that are currently being used in the Fox Insight data set, described below; the DQSr was based on a simulated (random) dataset representing each of these PROs structurally.

In addition to comparing DQSo to DQSr, we also investigated whether or not a fatigue effect was present in the FI data for these instruments. To accomplish this we incorporated the order in which these PROs are administered; if a fatigue effect is present in the response patterns, then the proportion of actual FI responses should tend towards random as the respondent moves through the set of PROs (i.e., earlier should have less random responses and later should have more). By evaluating DQS change over the survey section order, we can determine if there is an overall fatigue effect; by examining DQS according to the length (number of items) of each PRO in the analysis, we can also determine if longer instruments lead to greater likelihood of random responses (or responses that are closer to random than is expected). We also recovered observed Cronbach's alpha statistics to determine whether or not internal consistency, or psychometric characteristics, could predict the data quality for that instrument.

4. Results

DQSo scores are summarized in Table 2 as the average DQSo for that instrument (i.e., collapsing across participant). DQSr scores are summarized as the average for the same number of random respondents for that instrument. Cronbach's alpha was computed based on the standard formula (see Cronbach 1951), and the 95th percentile of DQSr was computed as described in Section 2.

Table 2: DQS results per instrument

	Instrument characteristics:				***Data Quality Scores:***		
Order	*Section*	*# Questions*	*Cronbach's Alpha*	*N*	***Average DQSo***	***Average DQSr***	***% below DQSr 95th percentile threshold***
1	PDQ8	8	0.69	9095	0.09	-0.02	64%
2	PASE	20	0.56	10726	0.17	-0.11	18%
3	EQ5D	5	0.52	10776	0.12	-0.09	66%
4	NUPDRS2	14	0.30	8323	0.12	-0.07	46%
5	Offsurvey	22	0.73	1836	0.18	-0.12	13%
6	PDAQ15	15	0.91	8179	0.19	0.00	36%
7	GDS	15	0.88	10150	0.15	0.02	65%

Table 2 shows that Cronbach's alpha, DQSo and DQSr were unrelated to the length of the instrument. The proportion of respondents whose DQSo scores fell below what would be considered a cutoff (threshold) for data that is unacceptably or unexpectedly similar to random responses ranges from a low of 13% (for the "OFF survey") to a high of 66% (for the EQ5D, a quality of life survey).

Figure 2 shows the data from Table 2 graphically, in the order in which the surveys were administered.

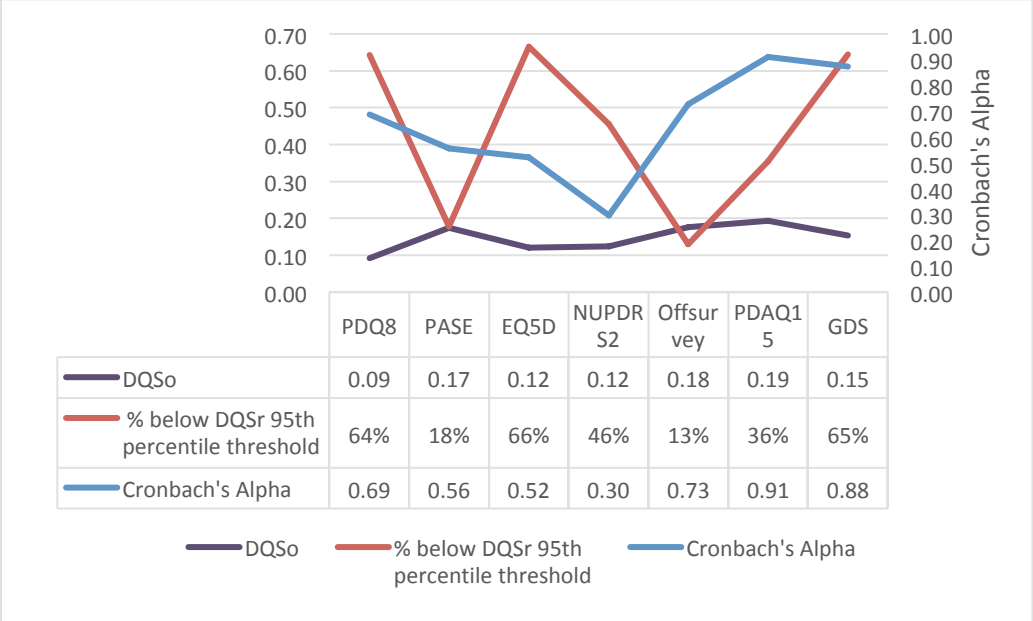

Figure 2: Comparison of DQSo with order of instruments (X axis), observed Cronbach's alpha, and DQSr scores.

A fatigue effect would have most likely resulted in a decreasing mean DQSo when DQSo averages for each instrument were ordered according to the order of PROs as administered in the baseline visit. This was not observed (see Figure 2).

Figure 2 also shows the proportion of « bad » data, defined as the proportion of DQSo for each instrument that fell **below** the 95th percentile of that instrument's DQSr (i.e., random responses) was unexpectedly high for three of the seven PRO instruments. For these three (PDQ8, which reflects quality of life; EQ5D which also reflects quality of life; and GDS, which reflects geriatric depression and mood) 64%-66% of the DQSo scores were comparable to the DQSr –i.e., 2/3 of respondents contributed data that was not different from randomly generated responses on those instruments. This was observed in spite of high Cronbach's Alpha score (e.g. all three alphas are >.52). Thus, data quality cannot be predicted by the psychometric characterizations (e.g., Cronbach's alpha) of the given instruments.

5. Discussion

A new method has been presented and discussed that investigators can use before they execute their planned analyses of existing data sets. The simulated data provides a distribution of DQSr scores that represents essentially the worst-case scenario for these PROs, as it would for any survey or PRO that had even marginal coherence. Our results suggest that the method is not unduly influenced by the length of instruments or their internal consistency scores.

A challenge for the FI project is that response rates are very high for first section of the set of surveys, but then response rates drop off; moreover there is a monotonically increasing drop off in responsiveness over time. These analyses can begin to address the question of whether this drop off is due to fatigue, and/or whether responses are becoming less good over time (possibly due to duration of the visit/complexity of the surveys, effects of Parkinson's, etc.). If the drop off was due to fatigue, then average DQSo should decrease (i.e., get closer to random) with longer instruments, and should decrease further into the "visit" (set of instruments to be completed at one timepoint); this was not observed. The drop off could also be attributable to insufficient time to complete all instruments, or the longer ones. While DQSo cannot directly speak to this, indirect evidence of this attribution would be DQSo decreasing only for later instruments, not for earlier ones (irrespective of length); this was also not observed. It is possible that instrument design or content may not inspire respondent confidence, and as they complete different surveys that may be redundant (e.g., there are multiple assessments of quality of life) or that may reflect constructs that do not resonate with respondents (e.g., assessments for mood and quality of life, which may have more relevance for scientists than in patients' day to day lives). In this case, DQSo should show within-instrument clusters of problematic items; some instruments, irrespective of their length and position in the "visit" should exhibit greater DQS with potential item combinations explaining most of the "badness". This *was* observed. The DQSo analyses suggest that some of the FI instruments need to be reworded/revisited, possibly with patient-centered focus groups around what features of Parkinson's disease **they** are most interested in sharing, rather than what features the FI investigators are most interested in assessing with the FI instrument collection.

The method can be used to detect, quantify, and then plan or choose the method of addressing nonresponse bias, if it exists, in any dataset an investigator may choose. It can also be used to diagnose challenges that may arise in one's own dataset, specifically arising from a misalignment of patient and investigator perspectives on the relevance or resonance of the data being collected.

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.* 201 Jun-Aug;37(3-4):197-212. doi: 10.1177/0193841X13518534.

Bollen KA. (1990). Outlier screening and a distribution-free test for vanishing tetrads. *Sociological Methods and Research* 19:80-92.

Bollen KA, Ting K. (1993). Confirmatory tetrad analysis. *Sociological Methodology* 23: 147-175.

Conrady S & Jouffe L. (2015). *Bayesian networks and BayesiaLab: a practical introduction for researchers*.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334.

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.

Glymour C, Scheines R, Spirtes P & Kelly K. (1987). *Discovering Causal Structure*. Academic Press: San Diego, CA

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.

Neapolitan, RE. *Learning Bayesian Networks*. Prentice-Hall, Inc., Upper Saddle River, NJ.

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.

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. (1978). *Psychometric theory (2nd ed.).* New York: McGraw-Hill.

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

Pearl J, Russel S. (2000). "Bayesian networks" UCLA Cognitive Systems Laboratory, Technical Report (R-277), November 2000. In M.A. Arbib (Ed.), (2003). Handbook of Brain Theory and Neural Networks, Cambridge, MA: MIT Press, 157—160. Downloaded from http://bayes.cs.ucla.edu/csl_papers.html 19 July 2018.

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.

Spearman C. (1904). General intelligence objectively determined and measured. American Journal of Psychology 15: 201-293.

Spearman C. (1927). *The Abilities of Man*. New York, NY: Macmillan.

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.