1-1-2011

Development of an Objective Motor Score for Monitoring the Progression and Severity of Parkinson's Disease

Timothy W. Albers
Portland State University

Development of an Objective Motor Score for Monitoring the Progression and Severity of Parkinson's Disease

by

Timothy W. Albers

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Science
in
Electrical and Computer Engineering

Thesis Committee:
James McNames, Chair
Martin Siderius
Richard Tymerski

Portland State University

Abstract

This thesis describes the development of an objective motor score (OMS) of Parkinson's disease that utilizes the Quantitative Motor Assessment Tool (QMAT) developed through efforts by the Intel Corporation and the Kinetics Foundation. Parkinson's disease (PD) is a movement disorder which is a member of a group of neurodegenerative diseases marked by the depletion or impairment of dopamine-producing cells in the brain. Since PD is chronic and degenerative, treatments are intended to either improve the quality of life for sufferers by superficially treating symptoms or slow and ultimately reverse the progression of the disease. No blood test or biomarker exists, so current assessment of the disease relies on a subjective tool called the Unified Parkinson's disease rating scale (UPDRS) which is a coarse scale that requires costly clinical administration and is subject to rater bias.

The objective motor score described in this thesis exhibits excellent clinimetric properties, having demonstrated usability, validity, reliability, and responsiveness. It was calibrated to the motor section of the UPDRS, but in addition to high correlation with the motor UPDRS, it demonstrated an excellent ability to track deep brain stimulation treatment levels and to detect improvement in motor function of subjects due to dopaminergic treatment. With an excellent intraclass correlation coefficient, the OMS is a reliable measure and due to the objective nature of the test, it does not suffer from rater bias. Though these results come from the development phase, they suggest that confirmatory studies will firmly establish the excellent properties of the OMS.

While further studies are in motion to improve upon the sensitivity of the OMS by exploring metrics of voice recordings and paced tapping tests, the OMS presented here is a complete and usable tool for assessing the severity of PD-related symptoms. In conjunction with the QMAT, it is ready to be used in clinical trials, clinical practice, and even in the homes of patients who suffer from PD. This makes it an invaluable tool that could begin to replace the UPDRS for use in PD research, reducing costs and confounding factors in studies as well as extending their capabilities into the home.

Acknowledgments

The author would like to gratefully acknowledge Dr. Lars Holmstrom who led the effort on this project and always gave invaluable advice, Brent Casady who tirelessly served as coordinator for the biomedical signal processing lab, and Forest Kernan who patiently waded through unpleasant data management tasks early on in this project. Thanks to all of the members of the biomedical signal processing lab who helped support the work. Special thanks to Dr. James McNames for offering the opportunity to assist in this research and for his constant support. Finally, without Craig Kinnie who volunteered his time freely, enduring exhausting testing and retesting of our ideas, we could never have learned so much from this project.

This research was funded in part by a grant from the Kinetics Foundation.

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

Parkinson's disease (PD) affects between five-hundred-thousand and one-million Americans and perhaps four to five million people worldwide [1, 2]. The disease is primarily characterized by four symptoms: tremor, rigidity, bradykinesia, and postural instability. Though these symptoms present in varying degrees and combinations for different individuals, they are chronic and degenerative, progressively worsening over time [1].

Tremor, the most often recognized symptom, refers to the trembling that can affect the hands, arms, legs, and sometimes even the jaw and face. Rigidity, or stiffness, often affects the limbs and the trunk. Bradykinesia is the term used to refer to a general slowness of movement throughout the body. Finally, postural instability refers to impaired balance and/or coordination [1].

PD is a member of a larger group of neuromotor diseases marked by the progressive death of dopamine-producing cells in the brain. Dopamine is a neurotransmitter that plays an essential role in the basil ganglia motor loop which is a critical part of the brain's control of the movement of the body. In addition to this, dopamine is involved in parts of the brain that regulate emotion, motivation, and the feeling of pleasure. This leads to a large range of non-motor symptoms of PD including depression, anxiety, loss of energy, compulsive behavior, and many more. A root cause of this loss of brain cells has been elusive and, as of yet, research has failed to

produce a widely accepted explanation [1].

Though the exact number is very difficult to accurately determine, some studies estimate that there are between five-hundred-thousand and one-million PD sufferers in the United States alone, with about fifty-thousand cases diagnosed each year [1, 2]. Early stages do not necessarily reduce quality of life significantly, but as the disease progresses it becomes increasingly difficult to perform everyday tasks and eventually sufferers are completely invalid and bedridden.

A disease as common and debilitating as PD is a large burden to society due to both medical bills and loss of productivity. The per patient cost has been estimated at about $25,000 per year [2]. If the high estimate of one-million sufferers in the US is accurate, then the annual burden of the disease for the country is about twenty-five billion dollars [2, 3].

Especially in early stages, existing treatments can dramatically improve the quality of life. Two of the most common treatments are dopaminergic treatment and deep brain stimulation (DBS). Both of these require a great deal of clinical monitoring and adjustment. This can be difficult due to the lack of an efficient and responsive tool for monitoring progression and severity of symptoms.

Because there is currently no known biomarker or blood test for PD, it is diagnosed and monitored based on evaluation of the symptoms. Currently, studies involving PD rely heavily on subjective rating scales [4]. These scales are often chosen to be the primary outcome measures in clinical trials and thereby form the basis for deciding if particular treatments have a significant effect when administered to groups with PD. In addition to their use for measuring effects across subjects, the scores or subscores from these scales are often used by physicians to help determine treatment type or

methodology and dosing for individual subjects.

Until the internal processes relating to PD are well understood and can be directly and conveniently monitored, physicians and researchers will have to rely on indirect tests that make use of observations or measurements of the symptoms, most notably the motor symptoms. This has led to much analysis and development of the current standardized rating scales as well as an increase in research on objective measures that would make use of quantitative measurements of motor function utilizing electromechanical devices [5].

Clinimetrics studies the properties of both rating scales and objective motor scores. This is a science in development, rooted in the more established fields of psychometrics and test theory [6]. It is concerned with the systematic validation of tests to be used as neurological outcome measures. The minimum properties of such an outcome measure, whether it be a rating scale or an objective motor score, are usability, reliability, validity, and responsiveness [7]. A detailed discussion of these will be given in subsequent chapters.

The Unified Parkinson's Disease Rating Scale (UPDRS) has been called the "gold standard" for quantifying the degree of impairment caused by Parkinsonian symptoms [8]. It is the most widely used standardized scale for assessing Parkinsonism [9]. A 2004 Movement Disorders Society critique stated that the strengths of the UPDRS are wide utilization, application across the clinical spectrum of PD, nearly comprehensive coverage of motor symptoms, and clinimetric properties, including reliability and validity [8]. After thorough review of cited studies relating to the clinimetric properties of the UPDRS, the declaration that reliability and validity are among its strengths appears to be an overstatement. Though many studies specify statistics

that claim to demonstrate these properties, most often they are accompanied by severe ambiguities along with vague interpretations.

Endeavors have been made to utilize electromechanical devices in obtaining useful measures of PD. One of the most important studies with positive results came from the simple use of a common midi keyboard. Termed Quantitative Digitography (QDG), this method demonstrated an ability to distinguish between PD sufferers and controls as well as patients in the "on" versus "off" state of medication. Though the conclusions from this study were limited, they increased the hope that such an objective measure might be developed and set the stage for further research [10].

Later studies demonstrated correlation between QDG metrics and the UPDRS Part III score (Motor Section) as well a clear effect when comparing patients in "on"/"off" medication states and "on"/"off" deep brain stimulation states [11]. Other studies have made use of Timed Motor Tests and more specialized equipment such as Motus, or Quantitative Repetitive Wrist Pronation-Supination (qr-WPS) which have shown good results [12].

The Kinetics Foundation and Intel teamed together to create a self-contained test battery based on active research in the area of objective PD measures. Initially known as the At-Home Testing Device (AHTD), the hardware was renamed the Quantitative Motor Assessment Tool (QMAT) and again renamed OPDM-Dexterity. This device incorporated many tests that demonstrated promise in previous studies, including pegboard, keyboard, button tapping, reaction time, voice, and tremor watch tests.

Christopher Goetz at Rush University led a feasibility study. This effort primarily focused on proving that such a device could be successfully and easily used in the home of patients and that the data gathered could be made available for analysis.

Although the secondary goal of demonstrating validity of the test battery was initially less successful, the feasibility and usefulness of such a device for at-home use was clearly demonstrated [13].

As a consequence of this feasibility study, a cross-sectional study was created for the primary purpose of establishing an objective motor score based on results from a revised test battery implemented on the QMAT. The data obtained from that study are the primary object of analysis for this research. This effort established an objective motor score (OMS) utilizing an improved test battery on the QMAT. Initial analysis demonstrated exciting properties of this OMS which suggest that such a motor test could be an essential tool in the future research and treatment of PD. With a fixed model created using the current data, the next step of analysis will utilize completely independent data to clearly establish validity.

Although a validation phase is still necessary to firmly establish some of the clinimetrics, this work took great pains to create an OMS which is very likely to possess the necessary properties. The QMAT has shown usability in previous work and the current test battery further established this, with patients reporting that they are stimulated by the ability to track their own OMS results day by day. Analysis showed good face validity, since the components present in the model reflect those that showed promise in previous research [10]. High correlation with the UPDRS motor scores substantiated convergent validity. Small changes between test-retest scores demonstrated test reliability for both PD patients and controls. Most importantly, the effect size for "on"/"off" treatment and rank correlation with DBS power levels demonstrated excellent responsiveness, reinforcing hope for strong performance results on new data.

Chapter 2
Clinimetrics

The essential characteristics for an acceptable neurological outcome measure, such as a rating scale or an objective motor score for Parkinson's disease, are well documented. These are components of the relatively new area of research called clinimetrics, the development of which has relied heavily on the more established sciences of psychometrics and test theory. The essential characteristics, or properties, can be broken into four categories: usability, validity, reliability, and responsiveness [6, 7].

2.1 Usability

The term usability encapsulates the idea that an outcome measure cannot be prohibitively expensive, time consuming, uncomfortable, or bulky. These qualities are obviously subjective and depend on the context. A test that is considered usable in a clinical trial at a medical facility would not necessarily be usable in the home. The main point of specifying this essential characteristic is to remind test designers of the fact that, regardless of the extent to which the other characteristics are met, an outcome measure will only be valuable if it is practical and realizable. It also suggests a criteria by which this factor could be used to decide on an appropriate outcome measure from a set of similar choices.

2.2 Validity

The concept of validity is a broad one which is typically broken down into three more specific categories. These are content, criterion, and construct validity [6]. Although these are related and may overlap one another to some degree, each is established by unique processes of research.

Content validity is supported by logical arguments. One considers the question, "Based on what is known, is it logical that this would be a valid outcome measure?" The answer to this comes from extensive research of literature, expert opinion, qualitative patient interviews, and by examination of other measures of the same or similar concepts. A test should be designed with content based on previous research in related topics that have shown promise. It should also be reviewed by experts in the field. Features that are generally agreed upon should be included and features which lack a general consensus should be carefully studied and probably eliminated. Feedback from patients should not be overlooked. If there are similar tests which are used for the same or a similar concept, the overlap should be reflected by similarity of content. Together, these steps are necessary to produce a test that the community accepts and uses in practice.

Criterion validity refers to correlation of the proposed outcome measure with an existing gold standard. In contrast with content validity which is established by logical inspection, criterion validity requires a statistical relationship to be established through quantitative assessment. In many cases, especially in the area of clinimetrics, such a gold standard does not exist. In this case, one must rely more heavily on construct validity to demonstrate overall validity.

Construct validity can be further broken down into four categories: 1) convergent

validity, 2) discriminate validity, 3) group differences, and 4) hypothesis testing [7].

1. Convergent validity is similar to criterion; however, instead of requiring correlation with a gold standard, correlation with other measures which attempt to quantify or categorize the same or similar ideas must be established. If various measures are highly correlated, then it may be said that they are converging toward the same measure, establishing the fact that a single construct is being measured.

2. Discriminate validity requires that a measure not be highly correlated with something that it is not intended to measure. For example, if one is attempting to quantify bradykinesia, or slowness, as a symptom of PD, one would expect that this could be disambiguated from a difference in reaction time between sexes. In other words, a measure of bradykinetic impairment should not be highly correlated with gender.

3. Group difference refers to the ability of a test to demonstrate a difference between groups that are expected to have different levels for the construct of the test. For example, a measure of the degree of impairment in Parkinson's disease ought to distinguish controls from PD sufferers.

4. Hypothesis testing requires that a measure support theoretical hypotheses which are based on the construct that the test purports to measure. As an example, PD is universally agreed to be degenerative. A measure of motor function, therefore, should support the hypothesis that impairment is worse after two years than at a baseline date.

2.3 Reliability

For a test to have any value it must produce a consistent score when the underlying construct is held constant. This concept is given the title reliability [6]. Two aspects are generally considered. These are 1) internal consistency and 2) repeatability [7].

1. Internal consistency, or inter-item consistency, is the extent to which the various elements that make up a test are measuring a single construct. It can be thought of as the natural stability that the test has due to the structure of its content.

2. Repeatability refers to the quality of being able to produce a stable score when a test is repeated under fixed or comparable conditions. Although this can be decomposed in some detail, two important aspects are intrarater reliability and interrater reliability.

   (a) Intrarater reliability is the repeatability of scores given by a single rater or device.

   (b) Interrater reliability is the repeatability of scores given across raters or devices.

2.4 Responsiveness

Responsiveness is the ultimate aim of any outcome measure [7]. It refers to the use of a test to demonstrate significant differences. In PD, these could be differences between the "on" and "off" states of dopaminergic treatment. Researchers believe that deep brain stimulation (DBS) power level has a monotonic relationship with PD impairment. An excellent demonstration of responsiveness, therefore, would be

a monotonic relationship between the test score and the DBS power level. This responsiveness is necessary to demonstrate any gradual reduction in symptoms or slowing of degeneration.

2.5 Assessment of Clinimetrics

Usability does not have an accepted quantitative assessment methodology, and it is unlikely that a reasonable standard could ever be established. There is no quantitative assessment for content validity. It can only be confirmed by the consensus of experts in the field. Because of this, it should be a part of the development process for an outcome measure. Otherwise, much effort could be wasted in the development of a test that never gains any acceptance.

To quantitatively assess criterion validity, a correlation coefficient is often used. Pearson's R is used if the measures are considered to have an approximately linear relationship. Spearman's rank correlation is used if a monotonic, though not necessarily linear, relationship is considered more appropriate. Both of these statistics result in a number between negative one and one, where negative one indicates a strong inverse relationship, one indicates a strong direct relationship, and zero indicates no correlation at all.

Assessment of the group differences criteria for validity can be done using the concepts of sensitivity and specificity. These are binary operators that quantify a measure's ability to distinguish between two groups. Sensitivity, denoted here as ϕ, assesses the ability of a test to positively identify subjects with the disability and specificity, denoted here as θ, assesses the ability to correctly identify controls. These

are defined by the following two equations:

$$\phi = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}}$$

$$\theta = \frac{N_{\mathrm{TN}}}{N_{\mathrm{TN}} + N_{\mathrm{FP}}}$$

where N_{TP}, N_{FP}, N_{TN}, N_{FN} stand for the number of true positives, false positives, true negatives, and false negatives, respectively.

Ideally a measure's sensitivity and specificity would both be one. For these to be calculated, a threshold must be established for which scores falling below are identified as "positive" and scores above are identified as "negative".

Sensitivity and specificity may be combined to create a receiver operator characteristic curve (ROC curve). This is a plot of sensitivity versus one minus specificity (or, alternatively, the true positive rate versus the false positive rate). For a sample of PD subjects and controls, the threshold can be adjusted from the maximum score to the minimum score obtained using the measure on all of the subjects. As it is adjusted, the sensitivity will start out at a value of one and the specificity will start out at a value of zero. Ideally, as the threshold increases the specificity will rise to one before the sensitivity decreases. The ROC curve would be a flat line at one and the area under the curve (AUC) would be one. This is equivalent to having the highest control score below the lowest PD subject score. If, on the other hand, the scores of controls overlap the scores of PD subjects, the sensitivity will drop before the specificity rises to one, and the ROC curve will be somewhere below the flat line at one and the AUC would be less than one. If the scores for both controls and PD subjects are uniformly distributed across the same range, then the ROC curve will

be approximately a diagonal line from the point (0,0) to (1,1), and the AUC would approach one-half, indicating that the measure does not separate controls from PD subjects at all. Thus, the AUC is a useful quantitative assessment of a measure's ability to separate groups.

Unless the test is always perfect in distinguishing PD and controls, AUC is not independent of the sample that is taken. For example, if the sample of PD subjects has advanced untreated symptoms and the controls are young and healthy, one would expect high AUC, whereas if the PD subjects are in an "on" state and the controls are age-matched, one would expect more difficulty in distinguishing the two groups and therefore the AUC would be less than one. This is important to note when reviewing research and when reporting results.

Analysis of reliability can be framed in terms of the following equation [14]:

$$S_X^2 = S_T^2 + S_E^2$$

where S_X^2 is the variance of observed test scores, S_T^2 is the variance of the true scores, and S_E^2 is the variance of the error between the observed test scores and the true test scores. The square root of the error variance is often called the standard error of measurement [14]. True test scores can be thought of as the expected score or the mean score that would be obtained if the test was repeated many times without any learning. Note that this error variance accounts for all sources of error including characteristics of the test, the individual taking the test, and the test administrator.

Reliability can then be defined as:

$$r_{XX'} = \frac{S_T^2}{S_X^2}$$

In practice, Cronbach's alpha or the intraclass correlation coefficient are often used to estimate this reliability. Assessment of internal consistency is most often done using Cronbach's alpha. This can be expressed with the following equation:

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K}\sigma_i^2}{\sigma_{\text{total}}^2}\right)$$

Cronbach's alpha is given by α. K is the number of components, items, or questions on the test. The variance of the i^{th} component is σ_i^2 and the variance of the total score formed by summing the scores of all components is σ_{total}^2. Notice that if each of the components is totally independent, then $\sigma_{\text{total}}^2 = \sum_{i=1}^{K}\sigma_i^2$, and $\alpha = 0$. Thus, total independence corresponds with no internal consistency. However, if every component is identical, then $\sigma_{\text{total}}^2 = K^2\sigma_i^2$ and $\sum_{i=1}^{K}\sigma_i^2/\sigma_{\text{total}}^2 = 1/K$ and Cronbach's alpha takes on the value of 1. Perfect correlation, then, corresponds to perfect internal consistency as measured by Cronbach's alpha [15].

One important point to consider when reporting or interpreting Cronbach's alpha is that the standard deviation of total scores, which depends on the diversity of the group of subjects sampled, affects the estimate of internal consistency. For example, a test may be made up of five identical metrics or components with additive independent random error. Consider a group of subjects with identical impairments who take the test. Notice that the differences in the value of test metrics, both between components and between subjects, will only be due to the random error in the metrics. Thus,

the five metrics will appear to be independent variables, $\sigma^2_{\text{total}} = \sum_{i=1}^{K} \sigma_i^2$, and the estimate of Cronbach's alpha will take on the value of 0. On the other hand, consider a diverse group of subjects with a wide range of impairments taking the same test. The difference in scores between components will have the same standard deviation; however, the difference between subjects will reflect the differing impairment. Thus, the metrics will be correlated, $\sigma^2_{\text{total}} > \sum_{i=1}^{K} \sigma_i^2$, and the estimate of Cronbach's alpha will be higher than 0. This demonstrates that the value of Cronbach's alpha is dependent, not only on the properties of the test, but also on the properties of the sample group used to estimate internal consistency.

The statistic most often used for establishing intrarater and interrater consistency suffers from this same problem. The intraclass correlation coefficient (ICC), here denoted by ρ, is an estimate of the following ratio [16]:

$$\rho = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}}$$

where $\sigma^2_{\text{between}}$ is the variance between objects of measurement, or subjects, and σ^2_{within} is the variance within objects of measurement, that is variance between scores given by the same rater at different times or different raters. Notice that as the variance between subjects decreases and the variance within subjects is held constant, the ICC decreases.

Like Cronbach's alpha, the value of the ICC depends on the diversity of the sampled group and has limited meaning without the specification of that group. One article suggests that the closeness of the ICC of the motor UPDRS for a study involving advanced PD to that of a study involving early PD sufferers confirms that

the estimate is "correct" [17]. There are two necessary conditions required to make this comparison valid. First, the distribution of motor UPDRS scores for the early PD sufferers should have the same standard deviation as that of the subjects with advanced PD. Second, the standard deviation of scores given by different raters or by raters who repeated the tests at different times should be the same. A simpler statistic that would characterize the reliability of a test score is the standard deviation of scores for repeated tests.

Responsiveness can be expressed with a normalized effect size. The effect size of a treatment is represented by the following equation:

$$\Delta = \frac{\mu_1 - \mu_2}{\sigma}$$

where Δ is the normalized effect size, μ_1 is the mean "on" treatment, μ_2 is the mean "off" treatment, and σ is the standard deviation of the sampled population. Estimates of the standard deviation of the sampled population are made in various ways, often using a pooled form [18]. This assessment of responsiveness is useful for comparing medications' effects on a single sample, but it does not generalize to allow for comparisons of medications between studies and cannot be used to interpret changes in score for an individual.

Chapter 3
Review of Literature

3.1 History of Rating Scales

After the first careful description of PD by James Parkinson in 1817 [19], little progress in the study of the disease was made for some time. Treatments and symptoms were explored, but the underlying causes remained undetermined until Carlsson's studies in the 1950s. He uncovered the biochemical changes in the brain which led to the clinical practice of prescribing levodopa as the most efficacious treatment for PD in 1967. Since then, researchers have developed many more treatments. These include surgical procedures, brain implants, medications, and physical therapy [1, 5].

To facilitate exploration of treatments, researchers studied methods for quantifying the symptoms of PD. As early as the 1920s, doctors developed electrical and mechanical devices for monitoring motor function [20, 5]. At the same time, researchers making efforts to evaluate the effects of treatments developed more subjective clinical measurements. Some of these clinical scales remain in use, including Schwab and England's scale based on the estimate of percent ability of total activities of daily living. A score of 100% indicated normal ability and 0% represented total helplessness. Each 10% increment included a definition to help physicians place patients' scores. In 1967, Hoehn and Yahr introduced a scale intended to determine progression of PD. This was a coarse scale with five defined stages ranging from early unilateral

symptoms (symptoms occurring on one side) to symptoms causing inability to walk at the last stage [20].

Many centers of PD research developed scores that quantified various symptoms (e.g. bradykinesia, rigidity, posture, tremor, speech) with 3, 4, and 5 point scales. From these scales, subscores could be calculated by grouping subsections together and simply adding the scores. An overall score was calculated by a simple sum of all of the subscores. Though many variants existed, important examples include the Webster scale, the Columbia scale, and the UCLA scale [20]. Researchers primarily needed these scales to evaluate the results of levodopa trials.

With various tools available for assessing specific symptoms as well as global severity, different scores were used at different locations. Difficulty in comparing results from study to study arose. This led to the suggestion that important features of these existing scales be aggregated into a single rating scale which could be utilized by everyone. Efforts culminated in the Unified Parkinson's Disease Rating Scale (UPDRS), completed in 1987 and described in detail in an article by Stanley Fahn et al [20, 5].

3.2 Unified Parkinson's Disease Rating Scale

The Unified Parkinson's Disease Rating Scale (UPDRS) has become a primary tool utilized by physicians and specialists in clinical trials, clinical practice, and other research venues [9]. The most widely used section is Part III (Motor Examination) which is comprised of 27 subscores that provide an estimate of the severity of the key Parkinsonian motor symptoms. Because this scale plays such a significant role in PD research, its clinimetric properties have been extensively studied. Although there are

many good results, there are significant drawbacks and uncertainties which suggest that the UPDRS Part III should be improved or, eventually, replaced [8].

The Movement Disorders Society has recently sponsored the development of a revised version of the UPDRS called the MDS-UPDRS [21]. Some of the clinimetrics properties of the new scale have been tested. Researchers concluded that the results supported the validity of the scale as a replacement to the old UPDRS [22]. Since this scale was first tested near the end of 2008, most current PD research still utilizes the older scale. For this reason, the properties of the older scale will be discussed here.

3.2.1 Usability

In many ways the UPDRS must be credited for being quite usable as evidenced by widespread use [8]. The UPDRS requires no specialized devices. However, it requires administration by a highly trained clinician, and it can only be completed in person. Video allows for many parts to be scored, but elements such as the examination of rigidity require the subject to be present. This not only causes problems for administration, but also for the thorough testing of the reliability of the UPDRS.

3.2.2 Validity

Face validity was well established by the careful design of the UPDRS [20]. Satisfactory convergent validity with other measures of PD has been suggested but not quantified [23, 24, 25]. The construct validity of the UPDRS has been "suggested" by various factor analyses of the UPDRS [23, 24, 25]. Finally, the sensitivity of the scale is claimed based on the generalization that studies have demonstrated sensitiv-

ity to change in clinical status. A detailed description of difficulty in establishing an minimum clinically relevant difference is given without significant conclusions. Thus, the major strengths of the UPDRS, aside from widespread acceptance, are not well established [4].

One factor analysis revealed six clinically distinct impairments measured by the motor UPDRS. These included three factors of bradykinesia, one for axial/gate and two for the right and left sides, one factor for rigidity, and two factors for tremor, one measured rest tremor and one postural tremor. Together, these six factors explained 78% of the total variation in the motor UPDRS for 294 patients, with the factors of bradykinesia and rigidity alone accounting for 68% of the variation. The same study demonstrated a high degree of internal consistency in the motor UPDRS (Cronbach's alpha, 0.95). This suggests that the motor section of the UPDRS gives a good single measurement of overall impairment severity [24].

3.2.3 Reliability

The importance of interrater and intrarater reliability has been recognized by experts in the field of PD, but it is difficult to characterize and estimate the degree to which rating scales possess these properties. In spite of this, they are often cited as strengths of certain rating scales.

In a 2003 Movement Disorder Society Task Force review of the UPDRS, reliability and validity are cited as strengths. These claims are not well substantiated within the report. Interrater reliability is vaguely described as appearing "adequate" according to a few studies [23, 26, 27, 28]. However, a detailed look at these studies suggests that the claims are not clearly supported and are not quantifiable or useful in the analysis

of new results. The claim that intrarater reliability is a strength of the UPDRS is also poorly substantiated with the statement that two studies showed "low to medium intrarater reliability" [29, 28] and another demonstrated a high estimate of the ICC [30, 4].

Statistics such as minimum clinically relevant difference and intraclass correlation coefficient, though intended to communicate conceptually, actually screen important results from studies. As an analyst trying to extract the information contained in the score obtained for a particular rating scale and make inferences thereupon, a simple report of the expected deviation between raters and within raters would be much more valuable. These would provide a simple context for interpreting changes in scores.

Fortunately, even though they have not been directly reported, some estimates of the intrarater variability can be inferred from the results of a few analyses. From one we can estimate the median absolute difference between scores given by a single rater on repeated tests of 0.09×42 or about 3.8 points for patients in the "off" state (mean motor UPDRS 42, standard deviation of about 13.25, ICC of 0.90) on the UPDRS motor scale and 3.9 for patients in the "on" state (mean motor UPDRS 15.75, standard deviation of about 9.4, ICC of 0.89) [17]. The standard deviation within raters, σ_{within}, can also be estimated from the ICC, again shown as ρ, using the following formula:

$$\sigma_{\text{within}} \approx \sqrt{\sigma^2_{\text{between}} \times \frac{1-\rho}{\rho}}$$

In the formula $\sigma^2_{\text{between}}$ is the variance between subjects. The estimate of σwithin is called the standard error of measurement [14]. This yields values of about 4.5 for patients in the "off" state and about 3.5 for patients in the "on" state. The ICC can

be used in the same way from a 2002 work by Siderowf et al. [30] to get a standard deviation within raters for the UPDRS motor score of about 2.8. Note that this came from a study of subjects with relatively low motor scores having a mean of 17.8, a standard deviation of 8.4, and an ICC of 0.9. Finally, in a more recent study of a sample of PD subjects with a mean UPDRS motor score of 19, a standard deviation of 12, and ICC of 0.89, a value of 4.2 was obtained [31]. In one study this value is reported directly as 2.48 [32]. Taking these examples together, one may infer that the standard deviation of test scores when administering the UPDRS motor section is somewhere between 2.5 and 4.5 points.

3.2.4 Responsiveness

There are examples of double-blind, placebo-controlled, randomized studies that have used the UPDRS to demonstrate significant improvements with treatment as opposed to with placebo [33, 9, 34]. In double-blind studies, neither the patients nor the raters have access to treatment information. Treatment is generally randomized so that levels do not depend on factors such as age, gender, and impairment level. When the UPDRS motor subsection was used as an outcome measure, either a basic test of significant change between mean "on" versus "off" medication or comparison of change due to treatment versus placebo was made [33]. Results in the form of basic tests of significance cannot be compared well to results using other motor scores to evaluate the same treatments. They demonstrate validity more than responsiveness.

In a study of ELLDOPA treatment in early PD sufferers, researchers found that after 24 weeks subjects treated with the medication decreased their motor UPDRS by 26.2% versus a 4.0% increase (P ¡ 0.001) for those given placebo. In terms of

raw UPDRS scores, those given ELLDOPA decreased their scores by 4.6 points and those given placebo increased by 0.2 points [34]. The average person who received treatment improved by an amount just about equal to the upper estimate of the standard error of measurement described in Section 3.2.3. This means that for an average individual the effect of levodopa could not be detected by the UPDRS.

3.3 Objective Measurement Tools

In the past several decades many researchers have focused on the development of objective measures that circumvent the necessity for human raters to make a judgment call when estimating the level of impairment for subjects suffering from PD. Many attempts quantify differences in levels of impairment that are too subtle for a human rater to distinguish. Countless examples utilize pegboards [35, 36, 37, 38, 39], button tapping devices [40, 41, 42, 38, 39], torque and strain devices [43, 44, 12], keyboards [45, 10, 43, 11, 44, 12], voice recorders, and accelerometers [41, 42, 12]. With these objective devices, measures have been developed in an effort to enhance or even replace the clinical rating scales. Among these some of the most significant results have come from button tapping devices, keyboards, and pegboard tests.

3.3.1 Button Tapping

Button tapping tests generally consist of at least two buttons placed adjacent to each other. Two tasks are often defined. The first measures average movement time, or tapping rate, by simply recording the number of taps alternating between two buttons in a given period of time. The second measures reaction time by giving prompts at unpredictable intervals and recording the amount of time lapsed between the prompt

and the action of depressing the appropriate button following the prompt. In general the measures of movement time tend to be significantly slower in PD sufferers than in age-matched controls and often highly correlated with UPDRS motor scores.

One study compared the tapping speed of normal control subjects to that of patients with PD. Two manual counters placed 20 cm apart tallied the number of complete presses during a 1-minute trial. Differences between the averages of PD subjects and controls tapping rate were clearly demonstrated. Significant correlation between age and tapping rate was detected in normal controls. A learning effect was consistent between dominant and non-dominant hands across controls and subjects with PD [40].

A button tapping device was created to quantify button tapping pace, movement time, and reaction time. Additionally, the Purdue pegboard test and UPDRS were scored. Significant improvement in movement time and reaction time tests was shown when comparing pre/post pallidotomy results for both contralateral and ipsilateral hands [46].

The BRAIN TEST utilized a PC to implement an early description of an alternating finger tapping test [47]. This computerized finger tapping test used the "s" and ";" keys on a standard computer keyboard pressed alternately for a period of 60 seconds. Four metrics were evaluated, two of which demonstrated significant correlation with the motor section of the UPDRS. Kinesia, simply the number of key presses in the 60 second interval, performed the best with a Pearson correlation coefficient of $R = -0.69$ ($P \leq 0.001$) [45].

3.3.2 Midi Piano Keyboard

A series of important studies has shown that the electronic midi piano keyboard may be used as an effective objective measure of PD. In 2000 a group at the Department of Neurology and Neurological Sciences at the Stanford University Medical Center investigated the use of a midi piano keyboard connected to a PC as an objective measure of Parkinson's disease. The study included 16 subjects with idiopathic Parkinson's disease in the "on" and "off" state of dopaminergic medication and also compared results to a group of 11 age-matched controls. The test was named Quantitative Digitography (QDG) and was used in several more studies at the Stanford Medical Center.

The task was defined in a way that reflected the finger tapping task from the UPDRS. Subjects were instructed to alternately press 2 keys with the index and middle finger for 60 seconds with each hand. Blindfolds and headphones with white noise eliminated visual and auditory feedback. Using a custom software, researchers extracted 6 metrics from the key strikes of each test. These included the mean and coefficient of variation (CV) of 1) downward velocity of strike, 2) duration of strike, and 3) frequency of tapping.

Qualitative analysis suggested that normal subjects performed with a nearly constant velocity, duration of strike, and frequency of tapping throughout the 60 second test. They also performed without any apparent fatigue or hesitation and a rapid rate of approximately 4 taps per second per finger and consistently with an average CV of 9%. On the other hand, subjects with PD showed behaviors that could be correlated with fatigue, tremor, freezing, and festination.

The quantitative results showed enormous promise, with every kinematic variable

demonstrating a significant difference ($P \leq 0.05$) between the "on" and "off" states of dopaminergic treatment. Improvement was reported as the change in value from the "off" to the "on" state divided by the "off" state value expressed as a percentage. For the mean values, improvement was greatest for downward velocity, 196%, then duration of strike, 47%, and finally frequency of tapping, 41%. The coefficient of variation turned out to be an even more sensitive measure of improvement with the CV of velocity improving an average of 229%, strike duration of 88%, and tapping frequency of 68%.

In this group of subjects and age-matched controls a perfect classifier was identified to separate PD subjects in the "off" state of dopaminergic treatment from age-matched controls. No subjects with PD in the "off" state scored lower than 30% for CV of velocity, whereas no normal subject scored above the 30% threshold. This suggested that a 60 second test measuring the CV of velocity might be used to identify PD sufferers. This measure, however, did not apply to PD sufferers in the "on" state, though there was a significant difference in the average of frequency tapping between those subjects and the age-matched controls [10].

A second study in 2005 utilized the same midi device and protocol to compare QDG to the UPDRS motor test in evaluating improvement in fine motor control from medication and deep brain stimulation. Although the test included 60 seconds of repetitive alternating finger tapping (RAFT), the analysis used only the first 30 seconds. Again, six metrics were defined. These were revised from the original metrics used in the 2000 study of QDG. They included key strike velocity, log of the duration of finger strike, log of the interval between strikes, standard deviation of velocity, log of the CV of duration of finger strike, and log of the CV of the interval between

strikes.

The results showed that four of the six parameters studied were significantly correlated ($P \leq 0.05$) with the UPDRS motor score. The strongest correlation was for log of the CV of duration of strike ($R = 0.66$; $P \leq 0.001$) followed by key strike velocity ($R = -0.61$; $P \leq 0.001$). Log of CV of the interval between strikes and log of duration of the interval between strikes also correlated ($R = 0.56$; $P \leq 0.001$ and $R = 0.50$, $P \leq 0.01$, respectively).

These factors also showed significant improvement during medication. Velocity showed the most improvement from both dopaminergic and DBS treatment ($P \ll 0.001$). No patient had a worse velocity "on" medication than "off" and only one had a worse velocity on DBS than off. Log of the interval between strikes and log of the coefficient of variation of key strike duration both showed improvement as well ($P \leq 0.001$ and $P \leq 0.05$, respectively).

A linear model, defined as the QDG composite score, fit a linear combination of three of these parameters to the UPDRS motor score. The model included key strike velocity, log of the interval between strikes, and log of the coefficient of variation of the duration of strike. This model predicted the UPDRS motor score with a correlation of $r = 0.704$ ($P \ll 0.001$). As an example of the sensitivity of the QDG, the composite score demonstrated a significant difference between treatments ($P \leq 0.05$) even when the UPDRS motor score was unable to distinguish the two ($P = 0.08$) [11].

Researchers also investigated the difference between therapies using each of the six kinematic variables from the QDG and found significant differences in three out of six of the variables. They found no difference for velocity or standard deviation of velocity ($P = 0.87$ and $P = 0.42$, respectively) and no difference for interval between

strikes ($P = 0.08$). They did, however, find significant differences in the duration of strike and CV of the duration of strike as well as the CV of the interval between strikes ($P \leq 0.05$).

Though other factors seemed to be apparent in the raw data traces for both of these studies, both concluded that the kinematic variables extracted from the QDG mainly measured bradykinesia. Support for this conclusion came from high correlation of the metrics with the bradykinesia subscore of the motor UPDRS [44].

A recent study included mean velocity of the key strikes from the more affected side (MA) as a measure of fine motor bradykinesia in very early stage, untreated PD. When compared to the performance of the non-dominant (ND) hand of age-matched controls, a significant difference ($P = 0.001$) was once again found. This further established the key strike velocity as an important tool that may be used for distinguishing sufferers of PD from normal subjects [12].

3.3.3 Pegboard

Pegboard dexterity tests have also demonstrated strong performance as indicators of symptom severity in PD. They measure several aspects of PD symptoms including speed of movement, tremor, and ataxia, [39] relying on both speed and fine motor control in the upper extremity. One study found that using test scores from their version of a pegboard test as the primary outcome in a trial that evaluates dopaminergic treatment could theoretically reduce the number of patients needed to 57% of the number needed when using the motor UPDRS score [38].

The Purdue pegboard test is used in a broad range of applications in various forms. It consists of two parallel columns of holes separated by a few centimeters.

When prompted, subjects remove pegs from containers at the top of the board one at a time and insert them into the holes, starting from the top and moving toward the bottom. The test is typically scored as the total number of pegs placed in a given interval of time. Research has demonstrated that the test provides a measure of finger dexterity; however, some research cautions that care should be taken to ensure that scores are reliable [36]. In spite of this warning, this pegboard test and closely related variants have been studied as an objective measure of PD.

Research on unilateral pallidotomy in PD analyzed scores from the UPDRS, the Goetz dyskinesia scale, and the Purdue pegboard test. Results from the pegboard test were used to indicate that there is not significant improvement in contralateral bradykinesia after pallidotomy.

Kraus et al. studied a variant of the Purdue pegboard test in which subjects were instructed to move twenty-five pegs over a distance of 25 cm from one set of holes into a corresponding set of holes. The movement time required was measured and results were compared for PD and age-matched controls. A highly significant difference in scores was found ($P \leq 0.0005$) indicating that the test produced an effective measure of PD symptoms [39].

3.3.4 Test Batteries

One study in the early nineties created an entire system for evaluating the impairment of subjects with PD. This included devices for measuring muscle tone, tremor, reaction time, and movement time. Peripheral devices for making these measurements connected to a PC computer with custom software for providing calibration, testing, and data analysis. Analysts compared results from PD sufferers to those of

healthy age-matched controls and demonstrated significant differences for metrics of mean horizontal and vertical displacement of accelerometers attached to the hands as well as mean movement time between the buttons of a peripheral device. Significant differences were also demonstrated in measurements of muscle tone as measured by a torque sensing device attached to the wrist. Although researchers concluded that the test was valid for clinical use in measuring tremor, muscle tone, and bradykinesia, the test battery gained little traction [41].

3.3.5 At-Home Testing Device

Finally, in 2009, an article by Goetz et al. established the feasibility of a device developed in a highly collaborative effort. The at-home testing device (AHTD) was the first fully self-contained tool for assessing PD symptoms using tests involving pegboard, digitography, button tapping movement and reaction time, voice recordings, and accelerometers for measuring tremor. Beyond the development of the device itself, a system for collecting and organizing data from research at clinics and in homes was tested and validated.

Initially research focused on feasibility. Although validation of the device's responsiveness to symptoms of PD was a secondary outcome, this aspect of the devices was not established in the early research. Nonetheless, the capabilities of the devices were modeled after much research that had taken place in individual objective measures, and future validation was highly anticipated [13].

3.3.6 Contributions of This Work

The work described by this thesis began the task of validating the AHTD as a useful device for evaluating Parkinsonian symptoms. Furthermore, an objective motor score was created which is ready to be used in clinical practice and clinical trials. To accomplish these tasks, data from both the original feasibility study and from a new cross-sectional study were analyzed.

The first step was to define a set of metrics and to extract them from the raw data obtained from the studies. Clinimetric properties of these metrics were then analyzed through statistics and visualizations. As a starting point for creating an objective motor score, the design team decided on a linear model. Two reasons for this decision were to reduce variability in the model selection process and to ensure an interpretable model for evaluation by clinicians and experts in the field. By adjusting parameters in the model selection algorithm, multiple models were obtained, and a report was created which summarized the clinimetric properties of each model. In a final review of this report, experts selected a model to be used as the objective motor score.

From the design of the hardware platform, test batteries, and user interface to the selection of the final model, this effort is the most thorough and rigorous approach that has ever been taken to create an objective motor score. This OMS has been implemented on a server that makes results available to authorized people via the internet. Although there is plenty of room for improvement, the only task that remains for the OMS to be usable as a primary outcome measure in clinical trials is its full validation on new data.

Chapter 4
Methodology

4.1 Hardware Platform

The QMAT device, developed by Intel, the Kinetics Foundation, and a consortium of neurologists, is a self-contained testing apparatus (See Figure 4.1). Its dimensions are 10 inches by 6.5 inches by 2 inches and it weighs 10.2 pounds. A two-key keyboard allows assessment of digitography or finger movement. Two buttons placed 173 mm apart can record data from repetitive hand tapping. An eight peg pegboard test assesses fine motor control of the hands. In addition to these onboard devices there is a docking station for the Actiwatch Device (Cambridge Neurotechnology, Cambridge, UK), a microphone port for voice recording, and an external foot tapper records data from repetitive foot tapping tasks [13].

The upper panel contains an LCD screen for displaying instructions and examples of the test requirements as well as a speaker for audio. Data is saved directly to a USB storage device. Control buttons allow the user to access complete instructions and to carry out all of the tasks from the test battery. [13]

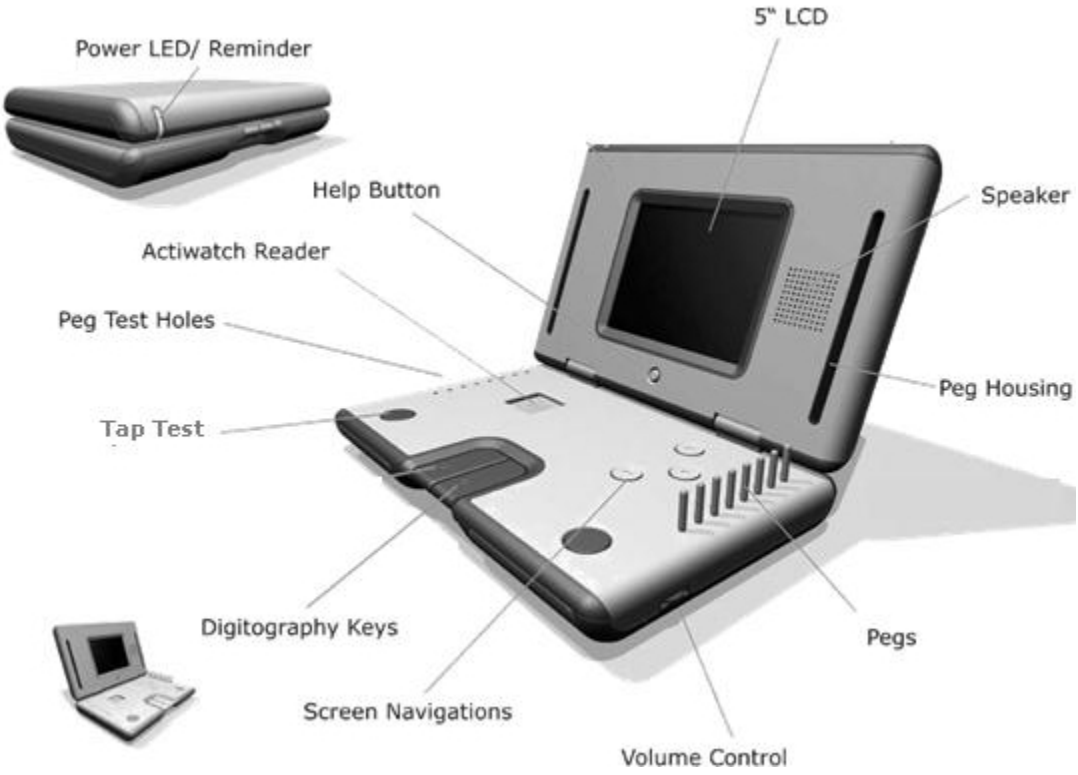


Figure 4.1: Illustration of the Quantitative Motor Assessment Tool.

Task	Dur. (s)	No.	Total (s)	Description
Finger Tapping	30	2 (LR)	60	
Digitography	30	2 (LR)	60	Trill test with first and second fingers. Called quantitative digitography (QDG). Eyes closed, ears muffled.
Reaction Time	30	2 (LR)	60	
Peg Board	30	6 (LR)	120	Each hand only once.
Sustained Phonation	10	6	60	Ahhh. Normal and loud volumes. Two repetitions of each. 48 kHz sample rate.
Story	50	2	100	Subject reads three passages. 48 kHz sample rate.
Total			460	7.7 minutes for test time (excluding directions).

Table 4.1: The above table summarizes all tasks included in the R1 test battery. Note that the total time listed above does not include the time required to set up the device, for the instructions to be given prior to each test, or for the data to be uploaded to the server.

4.2 Test Batteries

4.2.1 Release 1 Test Battery

Because research originally focused on feasibility, the first test battery included tasks that utilized each test type available on the QMAT device. Table 4.1 summarizes these tests. With the algorithms for processing the voice tests unavailable, the preliminary exploration of an OMS focused on metrics from the finger tapping, digitography, reaction time, and pegboard tests.

4.2.2 Release 2 Test Battery

The release 2 (R2) test battery was revised from the release 1 (R1) test battery with several goals in mind. While keeping the test as brief as possible, all tests believed to show any promise were to be included. Instructions for the test battery were to be sufficient for the test to be self-administered. In the previous study, subjects did not repeat the test and no controls were included to distinguish learning, boredom, and disease progression. Both of these features were added to the cross-sectional study.

Because experts suggested that foot tapping might be a sensitive measure of PD symptoms, engineers designed a foot pedal that took advantage of the microphone port as an input. Test battery designers then added a foot tapping task for each foot, augmenting the long list of motor tests and more completely covering the range of tasks evaluated with the motor UPDRS.

The motor test battery was augmented by the foot tapping task. Table 4.2 lists the names, time to complete, sides, and descriptions of each of the tasks in the R2 test battery. To shorten the length of the test, only one of each motor task was included.

Analysis of foot tapping data was left out because of trouble with the use of the foot tapping device, and analysis of the speech recordings was beyond the scope of this initial study. Therefore, this analysis only covers the four motor tasks utilizing the keyboard, pegboard, finger tapping, and reaction time tests of the QMAT device.

4.2.3 Release 3 Test Battery

In analysis of the foot tapping data, a lack of metrics with good performance combined with complaints received from patients in the clinic led to the decision to eliminate the foot tapping device completely. Patients also complained about the length of

Task	Dur. (s)	No.		Total (s)	Description
Foot Tapping	60	2	(LR)	120	Measured with a foot pedal. Based on [2, 3].
Finger Tapping	60	2	(LR)	120	
Digitography	60	2	(LR)	120	Trill test with first and second fingers. Called quantitative digitography (QDG). Eyes closed, ears muffled.
Reaction Time	30	2	(LR)	60	
Peg Board	30	2	(LR)	60	Each hand only once.
Sustained Phonation	10	4		40	Ahhh. Normal and loud volumes. Two repetitions of each. 48 kHz sample rate.
DDK speech	10	2		20	Puh-tuh-kuh. Two repetitions. 48 kHz sample rate.
Story	50	3		140	Subject reads three passages. 48 kHz sample rate.
Total				690	11.5 minutes for test time (excluding directions).

Table 4.2: The above table summarizes all tasks included in the R2 test battery. Note that the total time listed above does not include the time required to set up the device, for the instructions to be given prior to each test, or for the data to be uploaded to the server.

Task	Dur. (s)	No.	Total (s)	Description
Finger Tapping	30	2 (LR)	120	
Digitography	20	3 (LR)	120	Trill test with first and second fingers. Called quantitative digitography (QDG). Eyes closed, ears muffled.
Peg Board	30	2 (LR)	60	Each hand only once.
Sustained Phonation	10	4	40	Ahhh. Normal and loud volumes. Two repetitions of each. 48 kHz sample rate.
DDK speech	10	2	20	Puh-tuh-kuh. Two repetitions. 48 kHz sample rate.
Story	50	3	140	Subject reads three passages. 48 kHz sample rate.
Total			540	9 minutes for test time (excluding directions).

Table 4.3: The above table summarizes all tasks included in the R3 test battery. Note that the total time listed above does not include the time required to set up the device, for the instructions to be given prior to each test, or for the data to be uploaded to the server.

the digitography tasks. Analysis of the data showed that metrics performed as well on shorter segments of the data (20 to 30 seconds) as on the full length test (60 s), leading to the decision to reduce the length of the test to 20 seconds. In addition to shortening tests, repeated tests were added as a method for addressing the strong learning effect detected in the R2 data. All of this led to the third test battery, release 3 (R3).

4.3 Metric Definitions

A general framework was established for developing metrics for all of the tasks. This consisted of three steps.

1. The raw data recorded during the tests yielded a vector of actions. For example, the finger tapping tasks produced data with the following actions: right button depressed, right button released, left button depressed, left button released. Associated with each action was a time stamp, an action type, and a side.

2. Particular sequences of these actions were defined as events. Using the finger tapping tasks as an example, the sequence right button depressed, right button released, left button depressed, left button released, right button depressed was defined as a cycle event. The time elapsed from the first action to the last action was called the event duration and the time halfway between the start and finish was called the event time.

3. Statistics for each sequences of event durations were calculated and these were called metrics. Examples include the mean and standard deviation of the finger cycle durations.

In analysis of the R1 and R2 data, both parametric and non-parametric statistics were applied to these event durations and the resulting values were called metrics. These included multiple measures of central tendency, variability, and trend. As part of an effort to reduce the size of the metric pool, analysis in the R3 test battery included only parametric statistics. The following metrics were chosen:

- Mean of event durations.
- Standard deviation of event durations.
- Coefficient of variation of event durations.
- Linear trend slope of event durations versus event times.

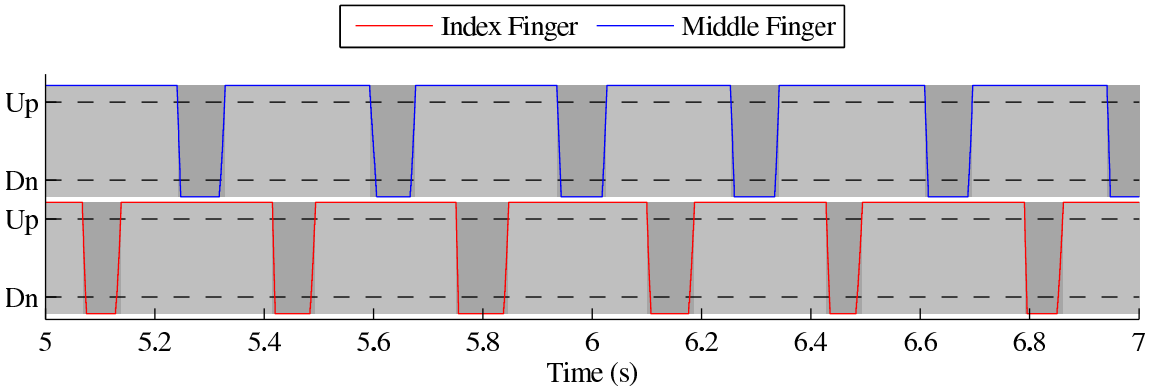


Figure 4.2: Example of raw data obtained from the QMAT keyboard test. The red and blue traces represent the relative key positions measured by the 20-stripe optical encoder of each key. The light and dark gray patches represent the "dwell" and "transition" events, respectively.

Because parametric and not robust statistics were used, a method of windowing called Winsorizing [48] was applied to reduce the effects of outliers. Finally, to help control the number of metric combinations, the metrics were calculated for both hands but only the metric value from the hand with the worst performance was included.

4.3.1 Task Types and Event Definitions

Keyboard

For the keyboard an action corresponded to the detection of a key movement by an optical encoder. For each action, the time stamp, key side, and direction of the key movement (up or down) was recorded. A full downstroke produced 20 detections by the optical encoder corresponding to 21 unique positions for each key. Using the sequence of these actions, each key location relative to these 21 positions was tracked through time.

Using these actions, the following events were defined:

- Transition duration: the time elapsed during a sequence of movements of a key up through the 3rd position and then back down through the 3rd position again.
- Dwell duration: the time elapsed during a sequence of movements of a key down through the 18th position, and back up again through the 18th position.
- Cycle duration: the time elapsed during a sequence of movements starting from the end of a transition event followed by a complete dwell event and finishing with another complete transition event.
- Downstroke velocity: the distance traveled divided by the time elapsed in the movement of a key from the 3rd position down through the 18th position.

Since the behavior of the index finger and the middle finger may more aptly show disability, these event values were recorded separately for each finger and were also combined into a set containing the values from both fingers.

Pegboard

Actions for the pegboard corresponded to the removal of a peg from a hole or the insertion of a peg into a hole. For each action the hole number, an indication of whether the action was an insertion or removal of a peg, and a time stamp were recorded. Based on these actions, the following events were defined:

- Transition duration: the time elapsed between the removal of a peg and the insertion of the peg on the opposite side.
- Dwell duration: the time elapsed between the insertion of a peg and the removal of the next peg.

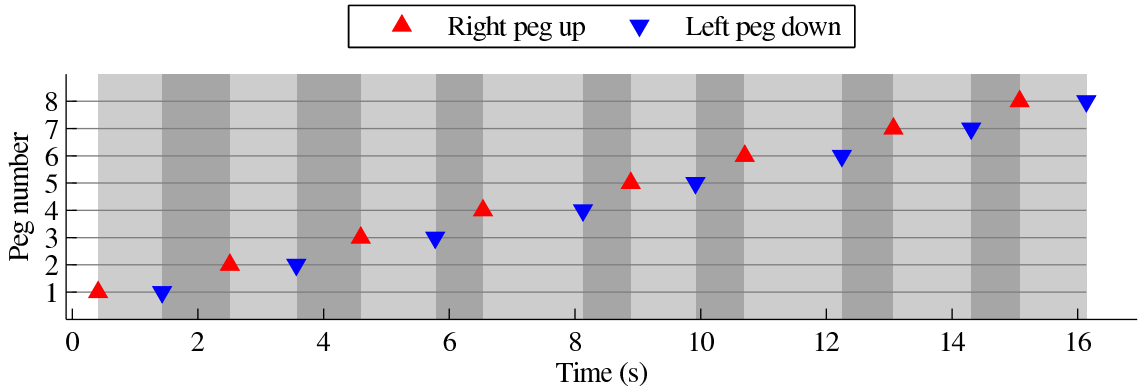


Figure 4.3: Example of raw data obtained from the QMAT pegboard test. The red and blue arrows represent the actions of removing or inserting a peg. The light and dark gray patches represent the "dwell" and "transition" events, respectively.

- Cycle duration: the sum of a dwell duration and the following transition duration.

Finger Tapping

The finger tapping actions recorded the depression of a button or the release of a button. For each action the button side, an indication of whether the button was released or depressed, and a time stamp were recorded. Based on these actions, the following events were defined:

- Transition duration: the time elapsed between the release of a button and the depression of the opposite button.
- Dwell duration: the time elapsed between the depression of a button and the release of that button.
- Cycle duration: the sum of a dwell duration and the following transition duration.

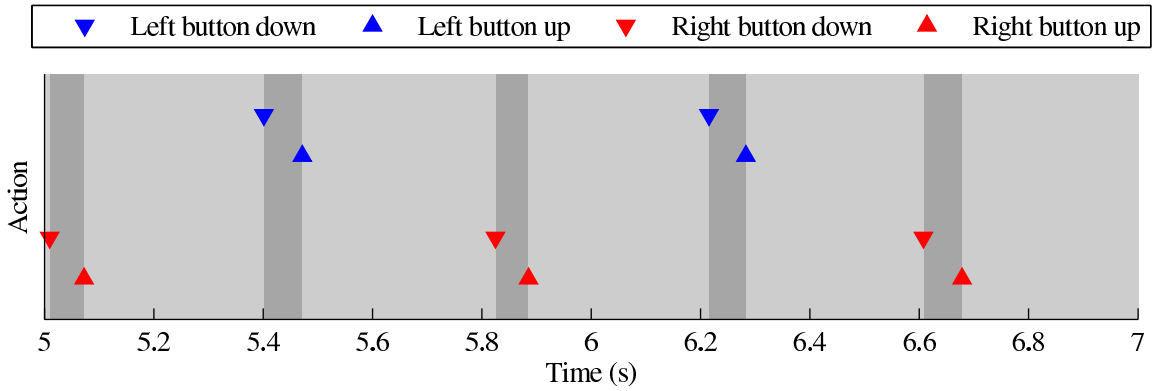


Figure 4.4: Example of raw data obtained from the QMAT finger tapping test. The red and blue arrows indicate actions of releasing or depressing the buttons on either side of the device. The light and dark gray patches represent the "dwell" and "transition" events, respectively.

Reaction Time

Actions recorded for the reaction time consisted of prompts, button depressions, and button releases. For each action, three values were recorded. The first value indicated whether the action was a prompt in the form of a beep, a left button action, or a right button action. The next indicated whether the button was being depressed or released. The last was a time stamp of the action. Based on these actions, the following events were defined:

- Transition duration: the time elapsed between the release of the button and the depression of the opposite button.
- Dwell duration: the time elapsed between a prompt and the release of the button.
- Cycle duration: the sum of a dwell duration and the following transition duration.

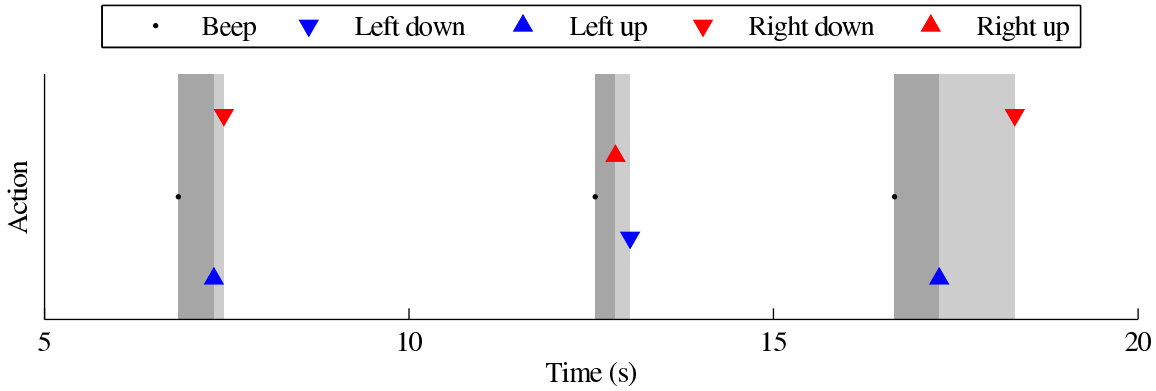


Figure 4.5: Example of raw data obtained from the QMAT reaction time test. The black dots represent the beep from the audio speaker prompting the user to begin moving to the opposite button. Red and blue arrows represent the actions of releasing or depressing the buttons on either side of the device. The light and dark gray patches represent the "dwell" and "transition" events, respectively.

4.3.2 Winsorizing

When appropriate, the metrics were Winsorized [48] to reduce the effects of possibly errant outlying values. This is a simple windowing process in which any value below a certain lower percentile is increased to the value of that percentile and any value above a certain upper percentile is decreased to the value of the upper percentile. In this case, a lower percentile of 3% and an upper percentile of 97% was chosen. Later, the substitution of log values for the metrics eliminated the need for Winsorizing.

4.3.3 Choice of Hand

Instead of having a separate metric for each hand, metrics from the two hands were combined into one by choosing the value from the side that showed the worst performance. This was done for two main reasons. First, combining hands cut the size of the metric pool from which a model was to be selected in half. For higher order

models, in our case three metrics, this cuts the number of possible models approximately in half for each metric added. More specifically, for a three metric model, the number of possible models is reduced by a factor of about $2 \times 2 \times 2$ or 8. By making the metric pool smaller, variability of the model selection process was reduced. Second, this reduces the confusion of defining and tracking a dominant hand and a most affected hand for each subject taking the test. This is important since many subjects are ambidextrous and some do not demonstrate stronger Parkinsonian symptoms in one side or the other.

4.4 Model Selection

4.4.1 Data

A data pair was defined as the data resulting from the completion of a QMAT test battery combined with the UPDRS motor score taken within two hours. Mismanagement of the R1 data resulted in the loss of a major portion of usable data pairs. Originally, there were 50 PD sufferers who were to complete the QMAT test battery three times in the clinic. Appointments were scheduled at baseline, after three months, and after 6 months. Ultimately, only 81 of the sessions were usable for model building.

The cross-sectional study was designed to quickly generate a large data set which was uploaded to a server for immediate analysis. In this way, test battery adjustments were determined and implemented as the study progressed. The R2 test battery included 108 data pairs taken by 57 PD sufferers and 19 age-matched controls. Data screening reduced this to 92 data pairs taken by 51 PD sufferers and 17 age-matched controls. 13 of the dropped data pairs resulted from incorrectly completed pegboard

tests and 3 resulted from incorrectly completed keyboard tests. Thus, 73 PD and 19 age-matched control data pairs were available for analysis (including data pairs obtained by subjects who repeated tests). Of these, 19 of the PD subjects took repeated tests that could be used to study OMS reliability.

Initially predictors were selected from a pool that included 12 finger tapping, 48 keyboard, 12 pegboard, and 8 reaction time metrics. Only metrics from the keyboard and pegboard tasks were selected by the stepwise process. Using all four tests resulted in the loss of more data due to errors encountered in the reaction time and finger tapping tests. Since the selection process ignored metrics from these tasks, the use of the data could be maximized by dropping them from the metric pool. The remaining pool included 48 keyboard and 12 pegboard metrics.

After an initial investigation of the R2 test battery data obtained in the cross-sectional study, changes to the test battery were implemented and R3 was created. A total of 302 R3 data pairs were collected. Due, presumably, to improvement to instructions and in event extraction algorithms, only 9 data pairs had to be dropped. Of these, 6 were due to unsatisfactorily completed keyboard tests and 3 were due to unsatisfactorily completed pegboard tests.

The remaining 293 data pairs were taken by 38 PD sufferers who repeated the test twice an hour apart, 23 controls who took the test each day for at least 4 days, 10 controls who took the test once, 6 PD subjects who took the test at varying levels of DBS power, 12 PD subjects who took the test "on"/"off" medication, and 8 other PD subjects who took the test once.

4.4.2 Selection Algorithm

For each release, the model was chosen by a simple stepwise selection process. In a series of steps, the model was augmented with a metric and the "leave one out" correlation with the UPDRS motor score was calculated. The metric with the highest value was tagged as a possible candidate for the model. Each metric was tested and the one contributing the highest correlation was chosen for the model. Next, the model was augmented with a second metric selected from the metrics remaining in the pool. Again, if the correlation increased, the metric was tagged as a possible candidate to be included in the model. The metric that increased the correlation of the model by the greatest margin was then kept in the model. This was repeated again until three metrics were included in the model. A model size of three metrics was chosen based on previous investigations, appearing to be the best choice considering the limited amount of data available.

4.4.3 Model Fitting

The selection algorithm produced a set of metrics that could be used to predict the UPDRS motor score with high "leave one out" correlation. These still needed to be combined into a single score. To create the objective motor score, a simple least-squares fit of the metrics to the motor UPDRS was used to estimate the parameters in the following linear model:

$$\text{Motor UPDRS} = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \omega$$

where X_i is the i^{th} metric, β_i is the i^{th} parameter and ω represents a zero mean Gaussian error term. After the parameters were estimated, an objective motor score

was defined by the following equation:

$$\text{OMS} = b_1 X_1 + b_2 X_2 + b_3 X_3$$

where the metrics are the same as above, but b_i represents the least-squares estimate of β_i from the model above.

4.4.4 Performance Measures

The feasibility study that produced the R1 data did not include age-matched controls, repeated tests, or "on"/"off" medication comparisons. Therefore, analysis was limited to evaluating models' correlation with the UPDRS motor score. Usability and convergent validity were evaluated, but reliability and responsiveness could not be considered.

More performance measures were available to be applied to the R2 OMS. Considerations included face validity, correlation with the UPDRS motor score, reliability, and the ability to distinguish PD from age-matched controls. Models were shown to clinicians and other experts to determine if the metrics used made sense. Correlation with the UPDRS motor score was evaluated based on the Pearson correlation coefficient. The ICC was used to compare models' reliability. Finally, the ability to distinguish between PD and age-matched controls was measured by the area under the ROC curve.

A more thorough examination of the R3 model was made possible by the diversity of the R3 data set. Performance measures were available for validating an objective motor score's usability, validity, reliability, and responsiveness. Beyond the usability already demonstrated by the feasibility study [13], total time for testing was consid-

ered to improve usability. Again, the model was reviewed by experts to determine if the metrics included in it made sense. Correlation with the UPDRS motor score was used to demonstrate convergent validity. The area under the ROC curve and comparison of means were used to evaluate validity based on the ability to separate distinct groups. ICC and standard deviation between repeated tests were used to demonstrate reliability. Effect size and comparison of mean "on" versus "off" OMS were used to evaluate responsiveness.

4.4.5 Statistical Methods

Establishing statistical significance and confidence intervals for the results in this report proved to be an unwieldy problem. Standard tests of significance and methods for creating confidence intervals exist for Pearson correlation coefficient, intraclass correlation coefficient, difference in means of both paired and unpaired samples, and Spearman's rank correlation coefficient. However, in general these methods require that the samples are taken at random from the distributions involved. In this case the entire data set is a random sample, but during the process of metric subset selection and model fitting, bias was introduced and the resulting objective motor scores cannot be considered random. Estimating the bias produced by the selection process is a difficult problem.

Analysis of the UPDRS has shown that there are several relatively independent factors which contribute to the overall score [25]. This reflects the fact that PD itself is a multifaceted disease with symptoms of tremor, rigidity, bradykinesia, and impaired balance and ambulation. Not only that, but the distribution of symptoms between subjects varies to a great degree, so that some suffer more from tremor and stiffness and others present stronger symptoms of bradykinesia [1].

In calculating a large number of metrics, the intent was to capture signals which were strongly related to as many of the symptoms as possible, and to find as small a subset of metrics that would be sensitive to those symptoms. Unfortunately, too many metrics increased the probability that a random sequence would be found that appeared to be highly correlated with the UPDRS.

These two issues express themselves, mathematically, as characteristics of the covariance matrix of the metrics augmented with the UPDRS motor scores. This covariance matrix can only be estimated based on the samples. If too few metrics are considered, then the multifaceted nature of the symptoms and their varying degrees between subjects will be missed, resulting in a high value for the standard error in estimating the UPDRS motor score. A covariance matrix with enough structure to express these complicated relationships would require more metrics. However, a large covariance matrix estimated using a relatively small number of samples has a high probability of containing unrealistically strong relationships to the UPDRS motor score.

Although some exploration of the structure of the covariance matrix and methods for estimating the bias by a resampling method called bootstrapping was attempted, the issue proved to be beyond the scope of this research. However, to give some sense of the levels of significance and the confidence bounds on the statistics, standard methods were applied.

Estimates of levels of significance (p-values) and a 95% confidence interval for Pearson correlation coefficient between the OMS and the UPDRS motor scores were based on the student-t test. The paired student-t test was also used for comparing the difference in means of paired tests, including repeated tests and tests taken "on"/"off"

of medication. For estimating the difference in the mean OMS for controls and PD subjects, the two-sample student-t test was used. 95% confidence interval for ICC was estimated based on analysis of variance [16]. To estimate the p-value of Spearman's rank correlation coefficient for OMS versus DBS power level, the exact permutation distribution was used since the sample sizes were small. Finally, to estimate the 95% confidence interval for the area under the ROC curve (AUC), a simple bootstrapping resampling method was used [49].

Chapter 5
R1 Analysis

5.1 Results

Using the data collected under the R1 test battery, approximately 1200 metrics were calculated. Note that there exist approximately 1200^3 or about 1.7 billion three metric models. These included metrics distinguished based on dominant, non-dominant, most affected, and least affected sides. Using this pool, the modified stepwise algorithm selected the following model.

$$\text{OMS}_{\text{R1}} = 2.63 + 13.56X_1 + 23.37X_2 + 14.96X_3$$

- X_1 - Mean of the middle finger cycle durations, right and left hands combined by maximum value, keyboard test
- X_2 - Coefficient of variation of the index finger dwell durations, right and left hands combined by maximum value, keyboard test
- X_3 - Mean of the transition durations, right and left hands combined by maximum value, pegboard test

This data did not allow for extensive performance measures. In fact, no controls or even repeated tests were available. This top model had a Pearson correlation coefficient of about 0.70.

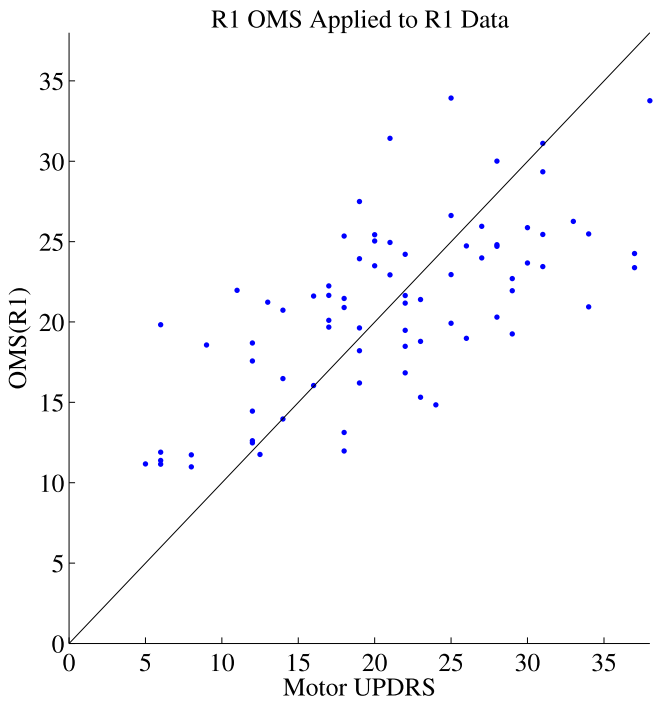


Figure 5.6: Scatter plot showing the correspondence between the R1 OMS and the motor UPDRS score for each of the R1 test sessions. Points falling on the diagonal line have an equal OMS and motor UPDRS score.

5.2 Discussion

No metrics came close to showing the same correlation with the motor UPDRS as the keyboard metrics, in particular the mean transition, dwell, and cycle durations. This chapter includes these R1 results in order to explain the careful analysis done on the R2 test battery.

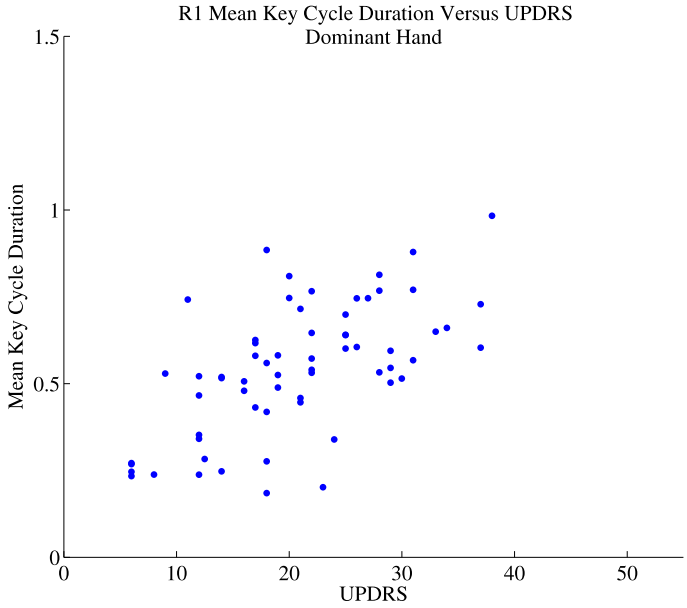


Figure 5.7: Mean key cycle duration demonstrated the most promise of any metrics in the R1 test battery. Much of the R1 OMS correlation with UPDRS comes from this single metric.

Chapter 6
R2 Analysis

6.1 Results

Using the data collected with the R2 test battery, the stepwise regression algorithm selected the following model.

$$\mathrm{OMS_{R2}} = -7.23 + 15.33X_1 + 16.43X_2 + 9.543X_3$$

- X_1 - Mean of the transition duration from the pegboard test, right and left hands combined by maximum impairment value.
- X_2 - Standard deviation of the middle finger transition durations from the keyboard test, right and left hands combined by maximum impairment value.
- X_3 - Coefficient of variation of the dwell duration from the pegboard test, right and left hands combined by maximum impairment value.

6.2 Correlation With UPDRS

The correlation with UPDRS motor scores was much lower for this R2 model and data set. The leave-one-out correlation value was only 0.52 and Pearson correlation coefficient was only 0.60. Most of this correlation comes from the first metric (R = 0.52). This only explains about 25% of the variation in UPDRS motor scores.

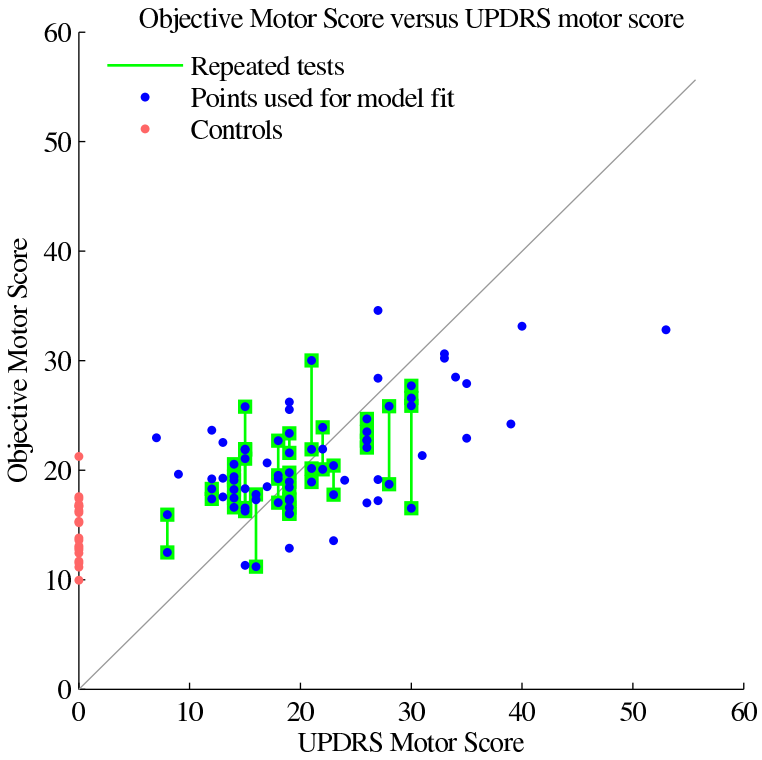


Figure 6.8: R2 objective motor score versus UPDRS motor score for PD sufferers. Points for controls are shown in red and points for PD are shown in blue. Tests taken one directly after the other are connected with a green line.

6.3 Ability to Distinguish PD Patients from Controls

Figure 6.8 shows the scatter plot of OMS versus UPDRS motor scores for subjects who took the R2 test battery. In this case, some subjects took the test back-to-back. The results from these two tests are connected with a green line. Also, the R2 data included age-matched controls. Their results are shown in red. These fall on the left side of the plot because their UPDRS motor scores were reported to be 0. With an AUC of 0.89, highlighted in green on Figure 6.9, this model showed a remarkable ability to distinguish controls from subjects with PD.

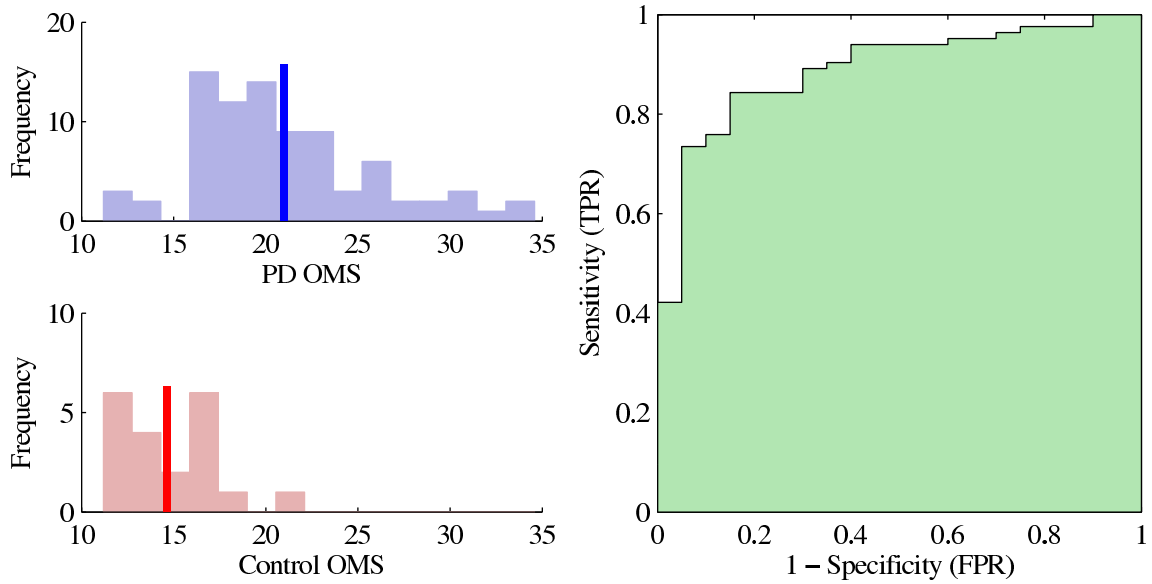


Figure 6.9: Top left is the histogram of PD patients' OMS. Bottom left is the histogram of control patients' OMS. Right side shows the receiver operator characteristic plot for distinguishing PD from control patients.

6.4 Repeated Tests

Though the AUC was high, this model demonstrated a very low ICC of only 0.31. In Figure 6.8 the long green lines show large offsets between scores on repeated tests. Figure 6.10 may explain this low ICC, showing a significant learning effect. On average, scores were reduced by 2.81 OMS points from test 1 to test 2.

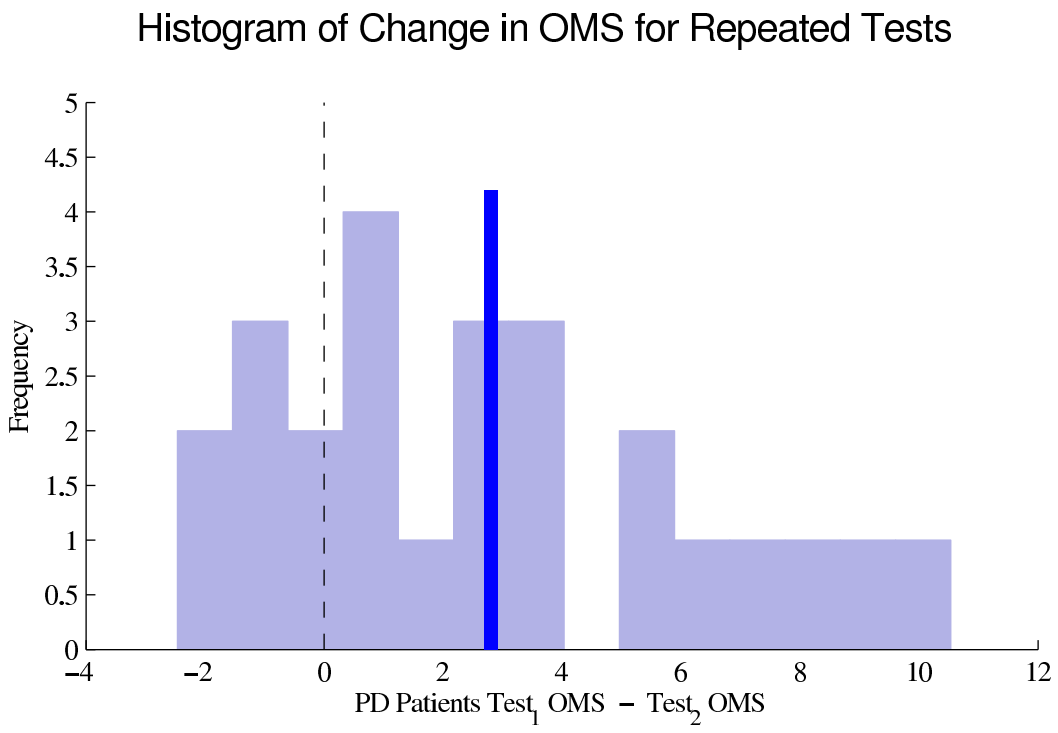


Figure 6.10: Difference of R2 objective motor score for tests repeated one after the other. Note that the histograms show the first test score minus the second test score so that a significantly positive mean may imply a learning effect in the test battery.

6.5 Discussion

Results for the model selection and performance evaluation of an R2 OMS model fell below that of the R1 model. Although in R1 the keyboard test produced the metrics that showed the clearest correlation with the motor UPDRS, for R2 the only metrics with the highest correlation came from the pegboard test. The need to understand this loss of performance in the keyboard test led to a careful investigation of the data obtained from the R2 test battery.

In Figure 6.11 a scatter plot of the mean key cycle duration versus motor UPDRS is shown. Although correlation is low, one may observe many points with high mean

key cycle duration and relatively low motor UPDRS scores. In general, patients with low motor UPDRS ought to be able to perform the test more quickly; they clearly did not. At the high end of the UPDRS, no subjects performed quickly. This results in a somewhat triangular shape in the scatter plot. All subjects with high impairment performed slowly, but subjects at lower impairments performed at a wide range of speeds.

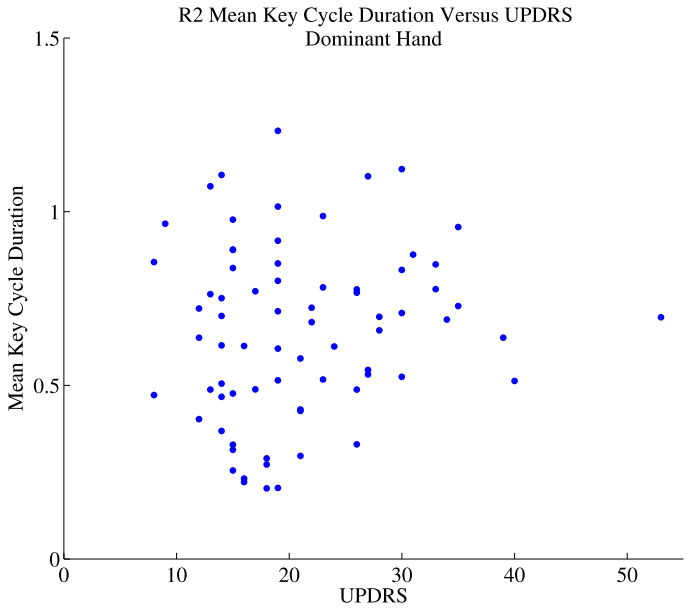
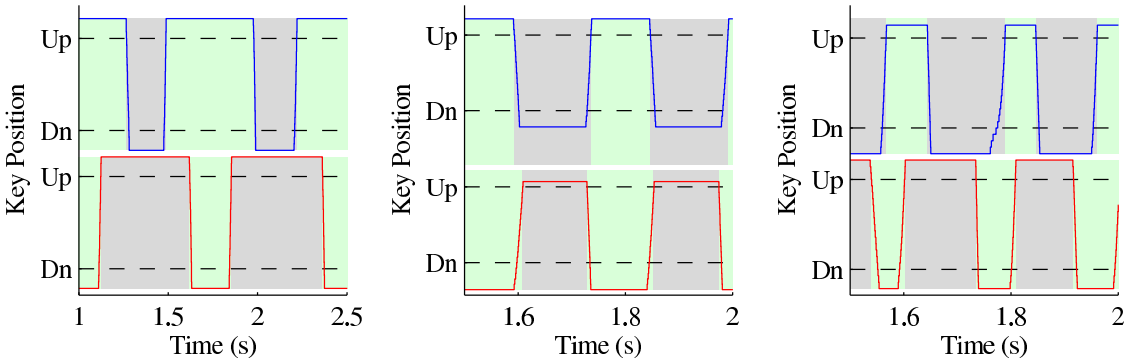


Figure 6.11: Mean keyboard cycle duration performed best among all metrics for the R1 test battery; however, the performance was lost in the R2 test battery.

Looking deeper into the raw data for the keyboard tests, investigation uncovered possibly important and somewhat unexpected characteristics of the tests. Figure 6.12 shows three unique strategies used by test takers to complete the keyboard test. The first, 6.12a, shows the pattern described in the QMAT instructions. One key is depressed and then fully released before the next key is depressed and fully released.

The second, 6.12b, demonstrates another, faster method, where the subject begins to depress the second key at the same time that the first key begins to raise. Though this may appear to be an intuitive motion, it does not match the instructions and gives the test taker an unfair advantage in speed over the person who correctly takes the test. A third strategy, shown in Figure 6.12c, appears to use a single motion to press both keys through a sort of rocking or galloping motion. This further increases the speed of the test taker and confounds the use of "mean key cycle duration" as a metric for impairment. Furthermore, some subjects actually changed strategies partway through the test as demonstrated by Figure 6.13.



(a) Correctly (b) "Teeter-Totter" (c) "Gallop"

Figure 6.12: (a) Raw data from a subject who took the test according to the instructions, depressing a key and completely releasing before depressing the next. (b) A different strategy where the subject began depressing a key at the same time that they began to release the other. Note that the time scales are different and the so called "teeter-totter" strategy is actually much faster. Finally, (c) shows a third strategy resembling a "gallop."

A second confounding factor in the keyboard test comes from the fact that different patients with PD present with different impairment symptoms. Thus, some may be fast but irregular, whereas some may be slow but very regular. These difference are

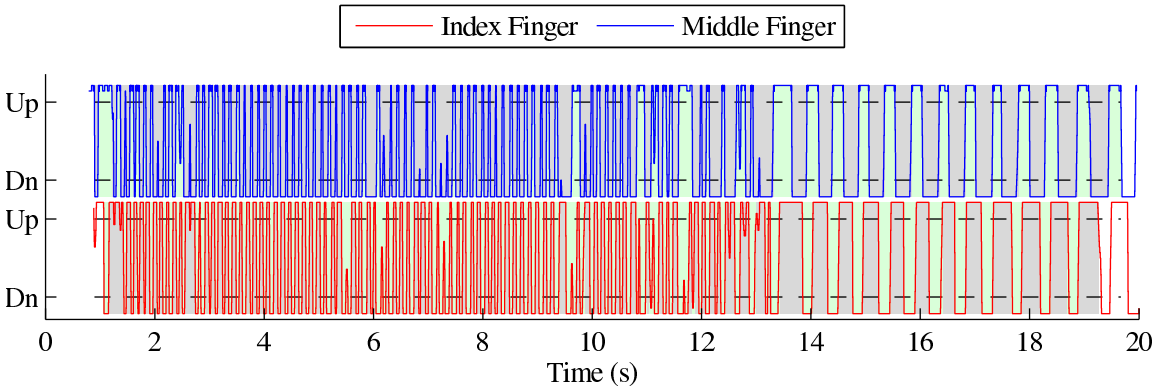


Figure 6.13: This raw keyboard data example shows that a subject changed strategies just over 13 seconds into the test.

demonstrated in Figures 6.14 and 6.15 with subjects who have about the same motor UPDRS score.

The discovery of several unexpected results while examining the keyboard data led to careful examination of the other tests from the R2 battery. Automatic reports with visualizations were created for the purpose of validating incoming data from each test as well as troubleshooting algorithms for automatically detecting events and calculating metrics. These tools produced key learnings for all of the test types.

Most data lost in the automatic event detection and metric calculation algorithms came from the pegboard. Visualizations allowed the source of these errant tests to be evaluated. Most commonly, patients removed two pegs at once and inserted both on the opposite side of the board. Also, patients commonly changed the order of the peg movements, and occasionally placed pegs into improper holes. Figure 6.16 clearly demonstrates this last behavior.

The finger tapping and reaction time tests utilized the same buttons. Visual analysis of the tests led to the realization that patients confused the two tests at

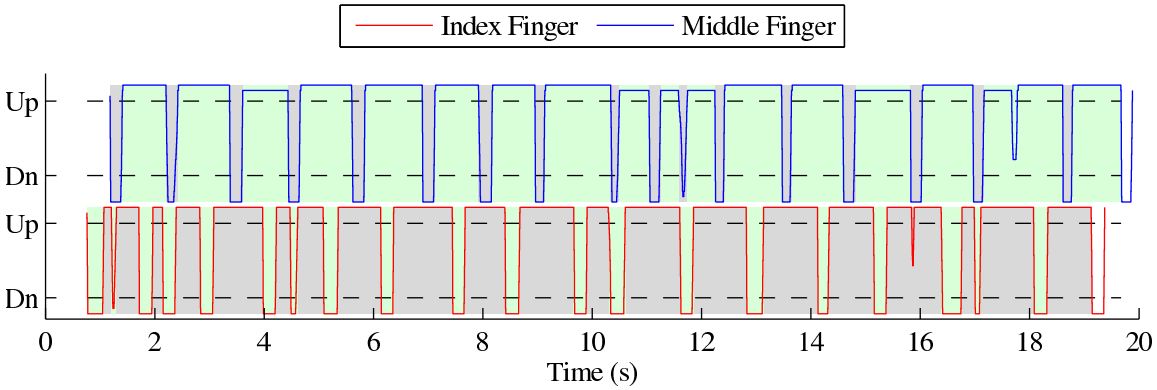


Figure 6.14: This subject's impairment was mainly characterized by extreme slowness of keystrokes.

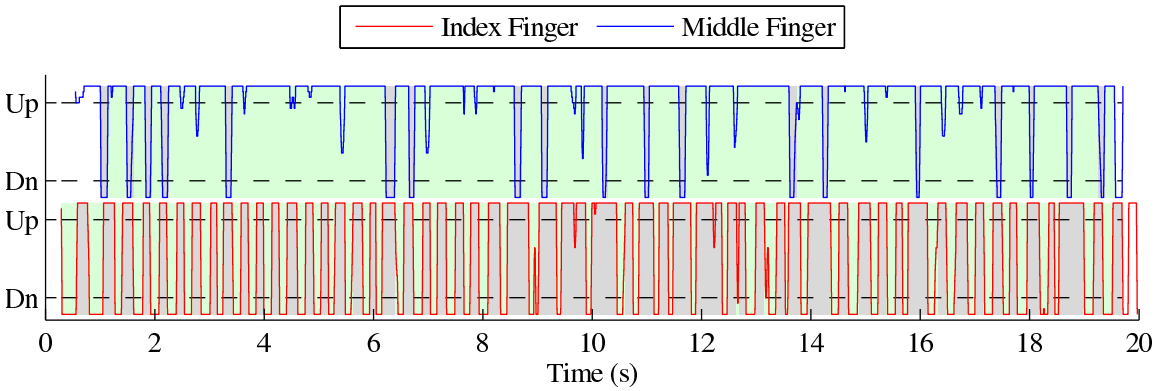


Figure 6.15: Though this subject was able to perform the keyboard test relatively quickly with their index finger, they were unable to consistently make full keystrokes with their middle finger.

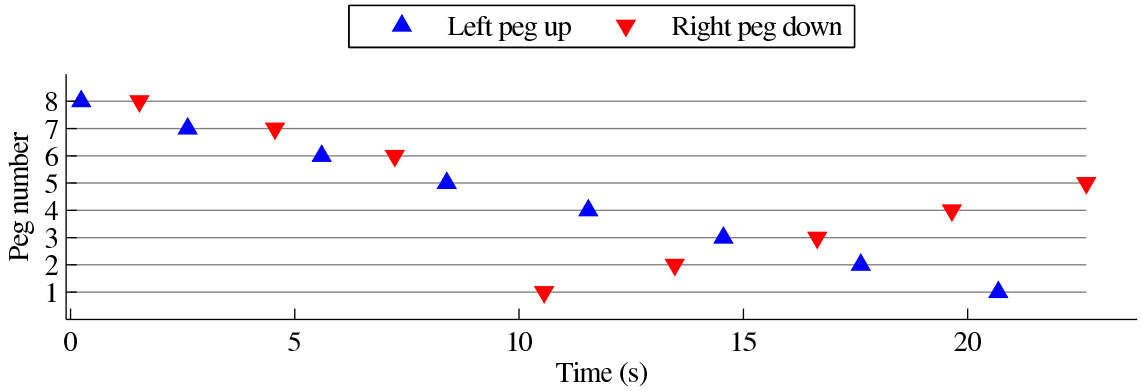


Figure 6.16: During this test the subject placed the first three pegs correctly, but the fourth peg was inserted, incorrectly, in the position of the first peg. This led to an entirely different ordering of the pegs and resulted in invalid peg data, making the OMS impossible to calculate.

times, exhibiting behavior during the finger tapping test that reflected the reaction time test and vice versa. Figures 6.17 and 6.18 demonstrate this behavior.

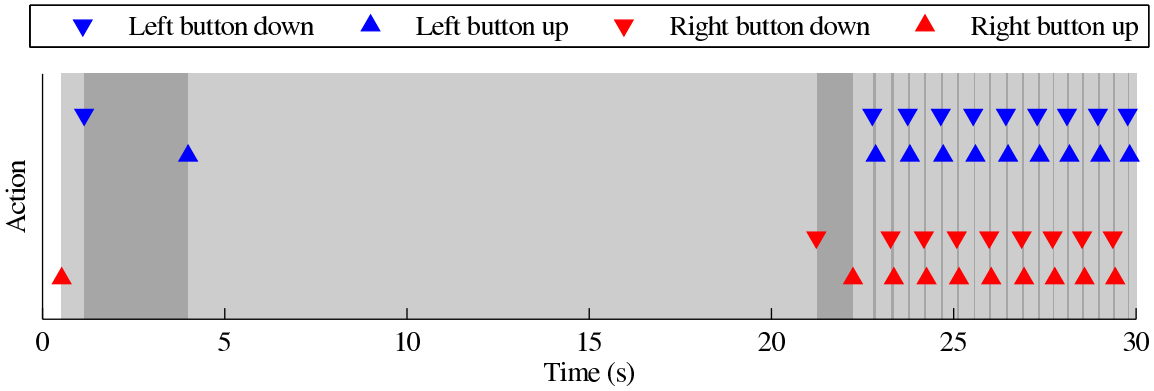


Figure 6.17: Though this data comes from a finger tapping test, it appears that the subject pressed a few buttons before pausing for nearly 20 seconds. It is possible that the subject confused the finger tapping test with the reaction time test, waiting for a prompt to continue pressing buttons.

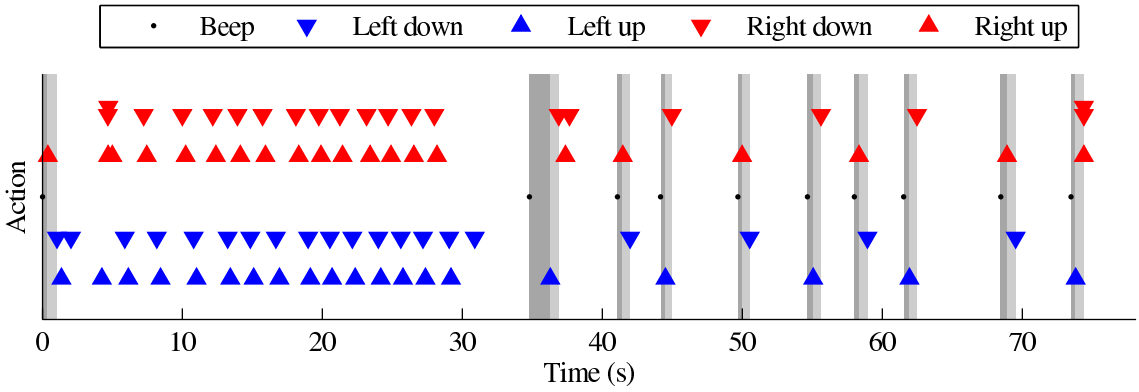


Figure 6.18: During the reaction time test this subject clearly did not wait for prompts from the speaker, but continued to alternate between buttons as though they were taking the finger tapping test. This continued for about 30 seconds before they stopped to wait for a prompt.

Chapter 7
R3 Analysis

7.1 Results

7.1.1 Model Description

The objective motor score is a simple weighted sum of three metrics which is represented in the following equation:

$$\mathrm{OMS}_{\mathrm{R3}} = -13.45X_1 + 16.87X_2 + 5.485X_3 + 82.2$$

where

X_1 Log of the Mean of the downstroke velocities from both keys, right and left hands combined by mean impairment value. From the keyboard test.

X_2 Log of the Mean of the cycle duration, right and left hands combined by mean impairment value. From the pegboard test.

X_3 Log of the Mean of the transition durations, right and left hands combined by mean impairment value. From the keyboard test.

The selection algorithm was based on optimization of the leave-one-out correlation with the UPDRS motor score. A least-squares fit to the motor UPDRS scores produced the final model parameters optimized to minimize the mean squared error across the available data set. Since most subjects took the test at least twice, the first and last test from each was used to fit the final model. The decision to use the first and last test was made in order to maximize the use of the available data without relying too heavily on any one subject. This final model resulted in a leave-one-out

correlation with motor UPDRS of 0.71 and a least-squares correlation of 0.73 (95% confidence interval: 0.63 to 0.80, $P \ll 0.0001$) as well as a mean absolute prediction error of about 6.21.

Figure 7.19 shows scatter plots of the objective motor score versus the UPDRS motor score. Blue points represent the data used to fit the final model. They include the first and last test from each subject with Parkinson's disease. If a subject only completed the test once, then that was also included as a blue point. When subjects took the test more than twice, the extra points were not used for fitting the model. Those extra points are shown in gray. In the few cases when a UPDRS motor score was given to a control, that data was included in the figure as a red point. Finally, pairs of points representing subjects who took the test back-to-back with no treatment are connected with green lines.

The initial pool did not include the log metrics. Visual review of diagnostic scatter plots of each of these metrics versus the motor UPDRS revealed nonlinear relationship, especially at higher values of impairment. Though mild across this data set, a loosely exponential relationship appeared in the duration as well as the velocity metrics. The addition of log values for both the keyboard test and the pegboard test resulted in better linear relationships with motor UPDRS and reduced the values of outliers, thereby eliminating the need for Winsorizing. With log values added to the metric pool, the model selection algorithm included all log values in the final model.

Table 7.4 shows significant statistics summarizing the performance of the R3 OMS. The first column shows the Pearson correlation with motor UPDRS. The second shows the mean absolute error. Note that this error should be compared to points on the motor UPDRS since the model is calibrated to that scale. The correlation between

	UPDRS		DBS	On/Off	Repeated Tests			Separation		
	R_P	MAE	R_S	Θ_{ES}	$\bar{s}$	$\overline{\Delta}_{12}$	ICC	Θ_{ROC}	$\bar{x}_{CS}$	$\bar{x}_{PD}$
OMS_{R3}	**0.73**	**6.21**	**-0.69**	**5.35**	**1.44**	**1.00**	**0.94**	**0.86**	**16.7**	**25.9**
$X_1(k)$	0.65	6.89	-0.83	-45.64	1.26	-0.10	0.93	0.67	22.5	25.8
$X_2(p)$	0.51	7.58	-0.76	1.42	1.38	1.41	0.88	0.90	18.5	25.8
$X_3(k)$	0.50	7.50	-0.64	2.40	1.45	0.49	0.85	0.73	21.4	26.1
# Pts.	120	120	54	24	76	76	76	97	33	64
# Subjs.	64	64	6	12	38	38	38	97	33	64

Table 7.4: Model statistics based on 64 patients with PD and 33 controls.

the R3 OMS and the motor UPDRS can be seen in the scatter plot of Figure 7.19.

Column three shows the average Spearman correlation (or rank correlation) with DBS power level for 6 subjects. This value is negative since the R3 OMS generally decreases as the power level approaches 100% of the optimal value. Figure 7.21 shows the results for each of the subjects. In all but one case, the Spearman correlation fell between -0.90 and -1.

The fourth column shows an estimate of the effect size for dopaminergic treatment. To make this calculation, the test/retest standard deviation was calculated from the subjects who repeated the test in identical clinical states. Here, the effect size reports the average change in R3 OMS from the "on" state to the "off" state divided by the test/retest standard deviation. In Figure 7.25 one can see that the change in OMS from the "on" state to the "off" state is positive in all but one case.

The next three columns summarize the results from subjects who repeated the test battery while in a constant impairment state. The average standard deviation between the first and second test is labeled $\bar{s}$ which is an estimate of the test-retest standard deviation. To measure the learning effect, the average change in OMS from test 1 to test 2 was calculated, labeled here as $\overline{\Delta}_{12}$. Figure 7.20 shows a histogram

of the change in OMS from test 1 to test 2 for all of the subjects who repeated the test. Finally, the intraclass correlation coefficient is shown in the column labeled ICC. In Figure 7.19, the green lines and boxes show the pairs of repeated tests used to calculate these statistics.

Under the heading "Separation," statistics summarize the ability of the OMS to distinguish PD patients from age-matched controls. The first, labeled Θ_{ROC}, shows the area under the receiver-operator characteristic curve (ROC) based on the first test of each control and each PD patient. Next, $\overline{\mathrm{x}}_{\mathrm{CS}}$ and $\overline{\mathrm{x}}_{\mathrm{PD}}$ show the average R3 OMS for the control and the PD patients' first test, respectively. Histograms of these scores, along with the actual ROC curve, are shown in Figure 7.24.

7.1.2 Correlation with UPDRS

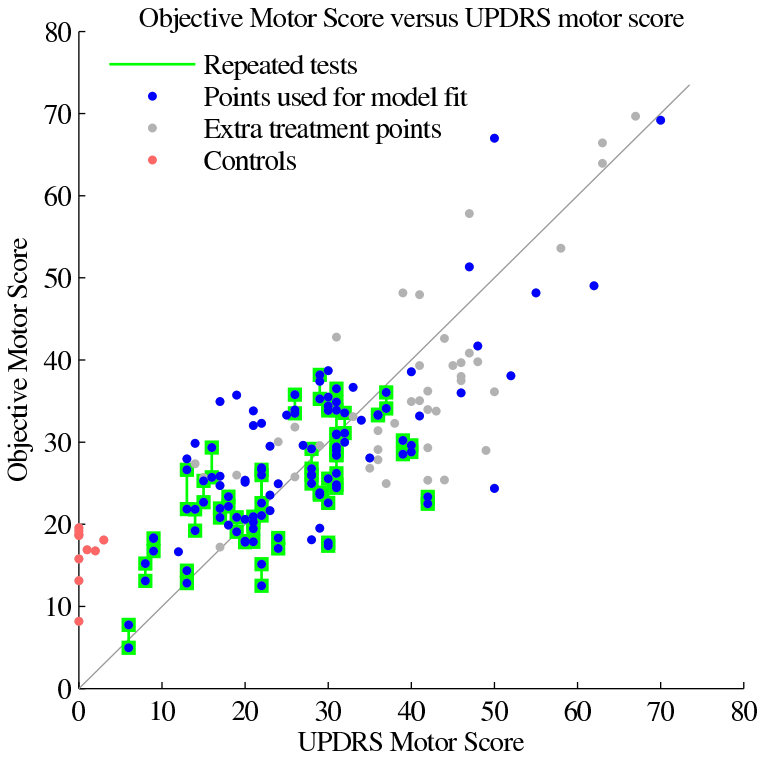


Figure 7.19: Scatter plot of R3 objective motor score versus UPDRS motor score. Blue points show the data used to fit the model. Grey points represent additional deep brain stimulation data used to help assess performance. Red points represent scores from controls. Tests taken one directly after the other are connected with a green line.

7.1.3 Repeated Tests

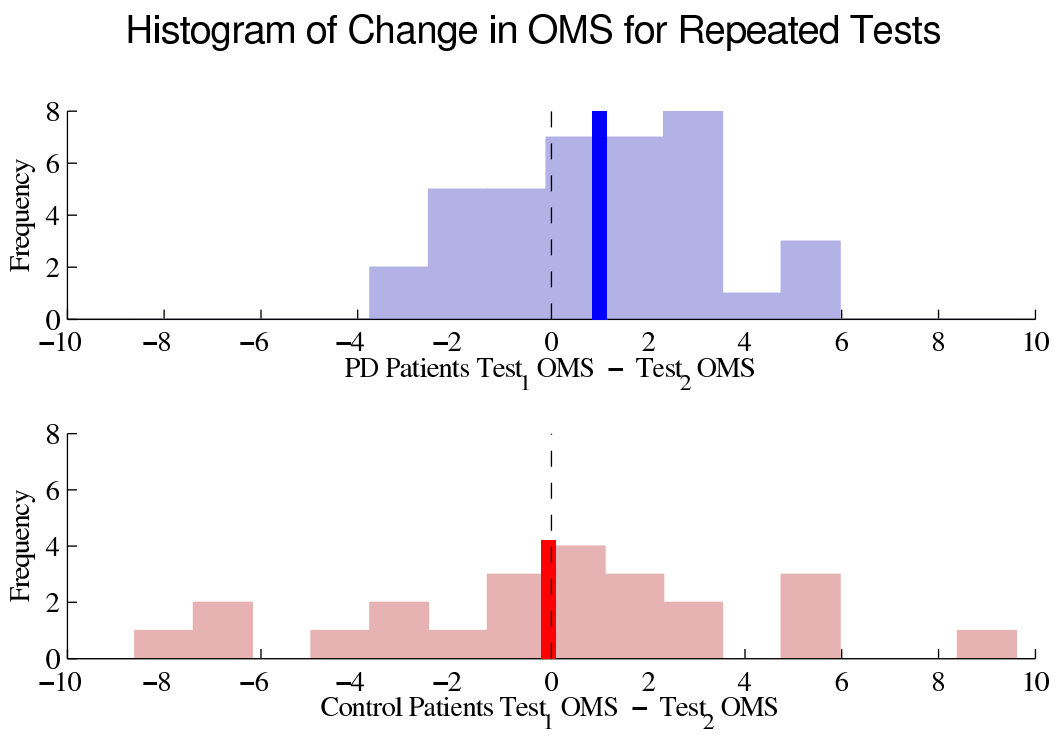


Figure 7.20: Difference of R3 objective motor score for tests repeated one after the other. Note that the histograms show the first test score minus the second test score so that a significantly positive mean may imply a learning effect in the test battery.

7.1.4 OMS Versus Treatment Levels

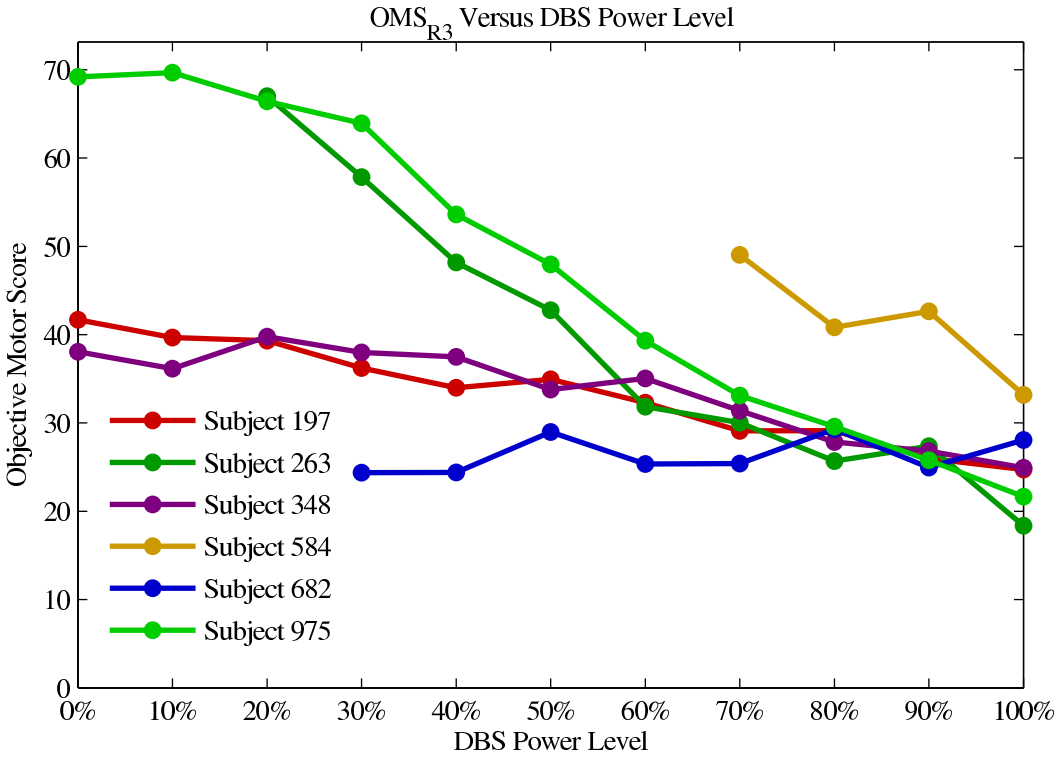


Figure 7.21: R3 objective motor score versus deep brain stimulation treatment level.

7.1.5 R3 OMS for Controls Across Five Days

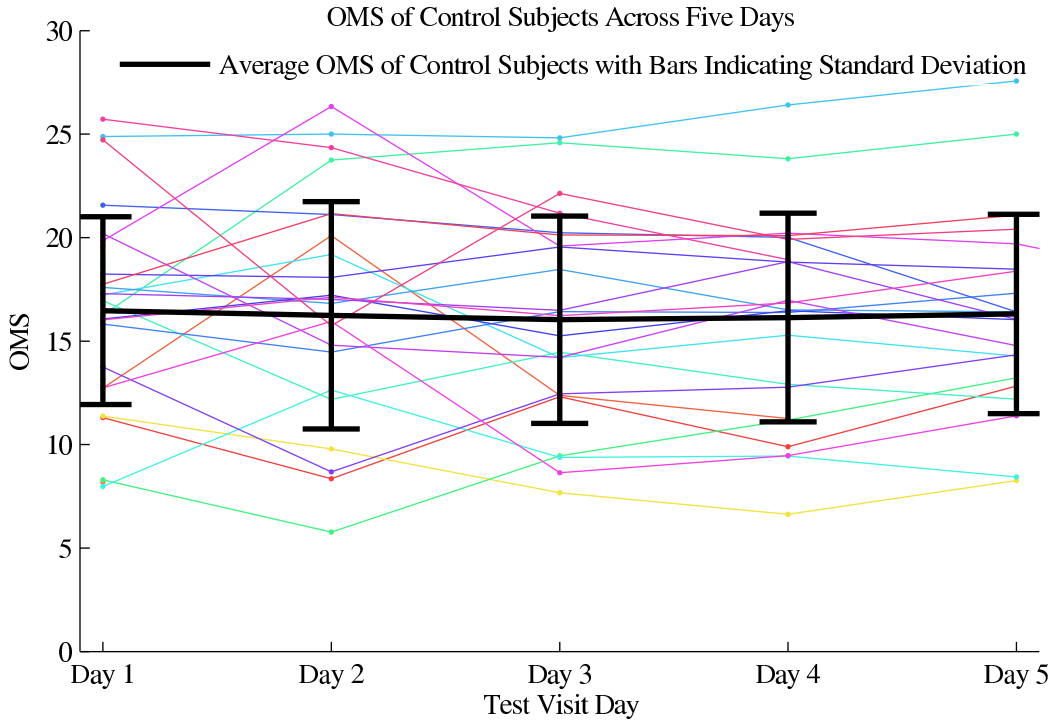


Figure 7.22: R3 objective motor score versus test day for controls who repeated the test battery once per day for a period of five days.

7.1.6 Changes of R3 OMS for Controls Across Five Days

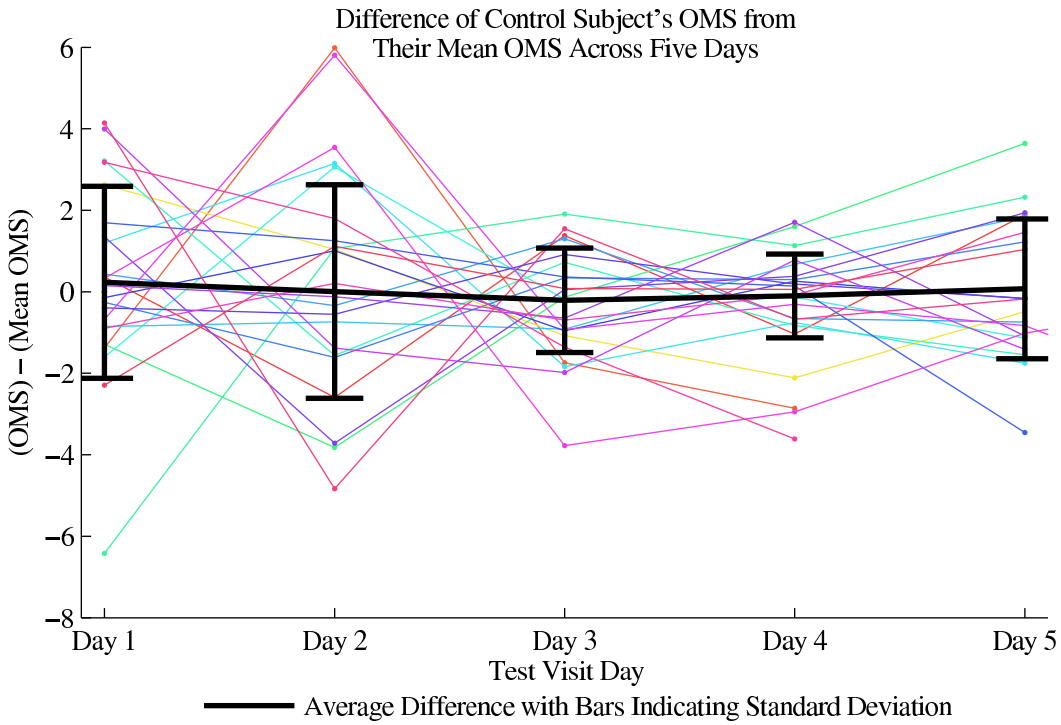


Figure 7.23: Difference from the average of R3 objective motor score versus test day for controls who repeated the test battery once per day for a period of five days.

7.1.7 Ability to Distinguish PD Patients from Controls

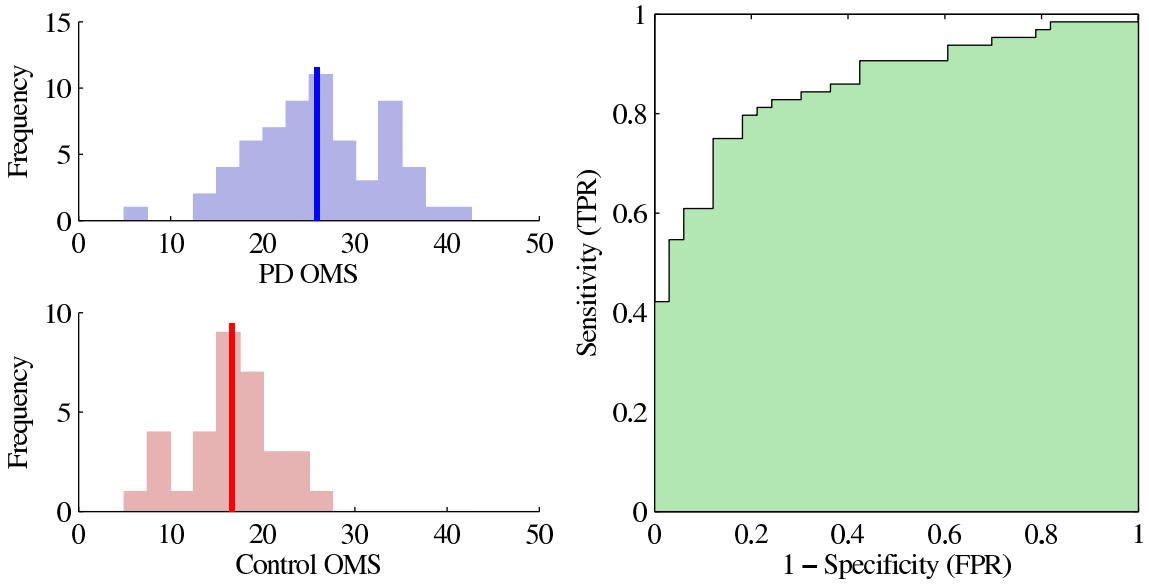


Figure 7.24: Top left is the histogram of PD patients' OMS. Bottom left is the histogram of control patients' OMS. Right side shows the receiver operator characteristic plot for distinguishing PD from control patients.

7.1.8 Change in R3 OMS in “on” versus “off” State

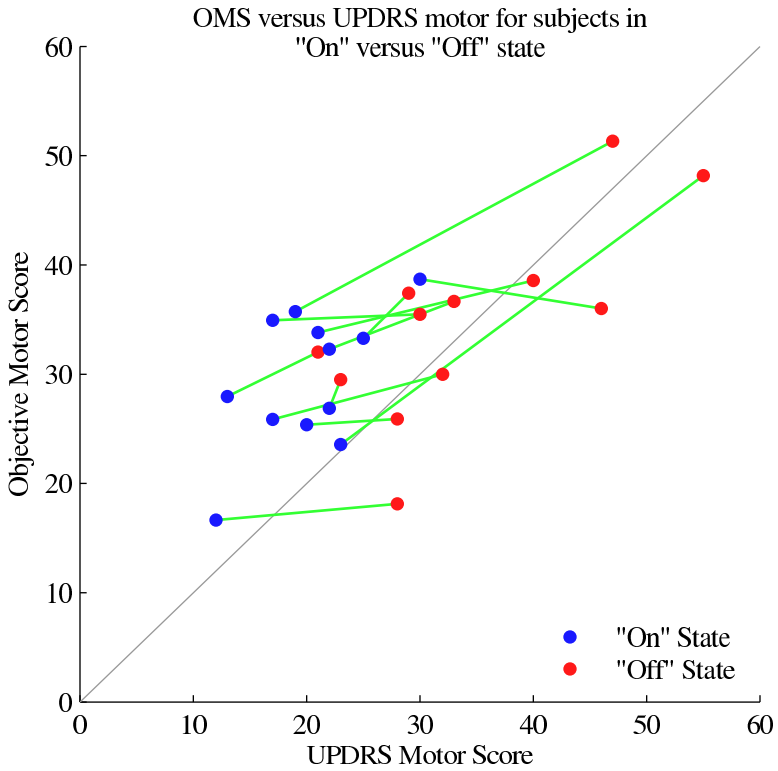


Figure 7.25: Scatter plot of R3 OMS versus UPDRS for subjects “on” versus “off” medication. Tests completed by the same subject are connected with a green line.

7.2 Discussion

These results from applying the R3 model to data from the cross-sectional study demonstrate excellent clinimetric properties in most of the important areas described in literature.

7.2.1 Usability

With the usability of the R1 test battery and the QMAT device firmly established by the feasibility study both in the clinical setting and at home, no further defense

is required to establish this property in the R3 test battery. In fact, the R3 OMS reduces the necessary tests to a very small subset of those in the R1 test battery. Only two task types are required to generate an objective motor score for patients and these can be completed in less than five minutes. Patients who have been given the opportunity to use the device on a daily basis have praised it, even saying that they looked forward to seeing their score each day.

7.2.2 Validity

Validity of the R3 OMS sits on firm ground. Content validity has been established by the fact that metrics in the model reflect those found in early experiments with a similar device [10]. Both downstroke velocity and mean cycle duration were determined to be strong metrics for distinguishing controls from PD patients in that study. In addition to this, the equation for calculating the OMS is clearly understandable, with simple metrics that have been reviewed and approved by experts in the field.

Although no gold standard exists for establishing criterian validity, construct validity is supported by every available measure. High correlation between the R3 OMS and the motor UPDRS score ($R = 0.73$, 95% confidence interval: 0.63 to 0.80, $P \ll 0.0001$) shows convergent validity. A significant difference of 9.17 (95% confidence interval: 6.42 to 11.92, $P \ll 0.0001$) OMS points between controls and PD patients demonstrates an ability to detect group differences. The AUC value of 0.86 (95% confidence interval: 0.81 to 0.91, $P < 0.0002$) further demonstrates this property. Hypothesis tests are supported for significant mean change of 5.34 points (95% confidence interval: 0.594 to 10.09, $P = 0.03$) in R3 OMS for patients in the "on" versus "off" state of dopaminergic treatment, with a mean "on" score of 29.6 and a

mean "off" score of 34.9. Similarly, a significant difference of 23.1 (95% confidence interval: 1.27 to 44.9, P = 0.041) was found between patients on 100% DBS versus 0% treatment, with respective mean values of 25.1 and 48.2.

7.2.3 Reliability

A high intraclass correlation coefficient of 0.94 (95% confidence interval: 0.89 to 0.97) suggests excellent reliability. More importantly, though, the very low standard deviation for repeated tests of about 1.44 compared to the value estimated for the UPDRS of 4 shows excellent reliability. A similar value of 1.0 for control subjects who repeated the tests each day for five days further establishes this important property.

7.2.4 Responsiveness

Results for both "on"/"off" medication and DBS treatment levels provide strong support for the most important clinimetric property, responsiveness. Of the nine subjects who repeated the test in the "on" and "off" states, eight showed a decrease in impairment going from the medicated state to the "off" state and the other showed only a very small increase. For the six subjects that repeated the test at a sequence of DBS treatment levels, four demonstrated an excellent monotonically decreasing relationship with a Spearman rank correlation coefficient between -0.9 and -1 ($P < 0.0001$) between R3 OMS and DBS treatment level. Though the other two subjects did not show significant correlation ($P = 0.33$ and $P = 0.20$), one decreased OMS with increased DBS stimulation and the other remained nearly flat.

In comparing the responsiveness of the OMS to the UPDRS one can see that the average change in OMS from the "on" to "off" state of levodopa treatment was 5.34

points and the average for the UPDRS was about 4.6 points [34]. These are close, but when considering the approximate standard deviation of repeated tests for the OMS of 1.44 points from section 7.1.1 versus 4.5 points for the UPDRS from section 3.2.3, it is clear that the change in OMS due to treatment is much more significant due to the reliability of the test. Of course, this must be validated by a study utilizing independent data.

7.2.5 Learning Effect

Compared to the strong learning effect detected in the R2 data, the R3 data showed a much reduced effect. This may be a consequence of both the improved instructions on the R3 test battery as well as the inclusion of practice tests, feedback, and repeated motor tasks. Though a significant learning effect was still detected, as apposed to the mean change in OMS from R2 of about 3 points, the change in OMS for subjects who took the R3 battery only improved by 1.0 point (95% confidence interval: 0.26 to 1.73; $P = 0.0093$).

Chapter 8
Conclusions

8.1 Summary of Contributions

In summary, this thesis introduced a new tool created to precisely and efficiently characterize the level of impairment of people suffering from Parkinson's disease. After describing the need and purpose for such a tool, a review of relevant literature was made to lay the groundwork for the analysis. This included a discussion of the current practice of using the UPDRS, related efforts to develop objective measures, and the tools used for analyzing clinimetric properties.

An initial methodology was developed for creating the objective motor score, and the results of applying this methodology to the R1 and R2 test battery were presented and discussed. When unexpected results were obtained, a deeper analysis was carried out and remedial adjustments were made to the test battery based on this analysis. The R3 test battery did indeed exhibit improved properties and based on this test battery, an objective motor score with good clinimetric properties was created.

8.2 Conclusions

The feasibility of the use of the QMAT device had been firmly established, but previous research had not demonstrated a practical method for using the data obtained to

quantify the impairment of subjects with PD. This research has taken the next step and helped to established a complete test battery along with a corresponding motor score that is ready for use, both in the home and clinical settings.

This objective motor score is a significant contribution because it is an efficient, precise, and simple to use tool for the assessment of impairment for people with Parkinson's disease. It is important because PD is an extremely widespread, costly, and debilitating disease. The objective motor score presented here may serve to make clinical trials of new treatments more efficient and effective. A major advantage that it holds over current practice of using clinical rating scales is its reliability. Results from this motor score will be directly comparable from study to study and year to year. In this way it can aid in the development of new treatments and hopefully someday a final cure for Parkinson's disease.

Though this research was investigatory in nature and could have failed to produce a valid measure, it should not be surprising that the results were successful. Many resources and much research went into the design of the hardware device. It is based mostly on earlier research that showed promise. For the first time, though, these tests have been compiled into a single test battery which yields a motor score that can be used for disease assessment, whether it be used for single measurements of impairment or for tracking impairment longitudinally.

8.3 Recommendations for Further Work

The next step in the process is to completely validate the OMS by applying it to a new independent data set and firmly establish its clinimetric properties. This could be completed with a comparable study to that of the new MDS-UPDRS [22]. Although

the OMS appears to possess good properties, and measures were taken to help ensure this, its validity cannot be proved without a follow-up study. In addition to this, the following actions, some already in progress, are recommended.

- New QMAT devices with an adjusted range of angular detection on the keys have been fabricated. Adjustments will have to be made to the current algorithms to allow the objective motor score to be calculated for these devices. Further analysis of the advantages of this change should also be carried out.

- A new test battery, Release 4, has been implemented and data is being collected at sites around the world. This test battery includes a paced-tapping test for the keyboard, similar to one that has shown promise in a previous study [50]. New algorithms for extracting metrics using the new test are currently under development.

- Since there is an identical keyboard and pegboard test in the R4 test battery as that in R3, the R4 data should be considered to potentially validate the current OMS.

- The analysis software could be improved. Interface to remedial tools, model selection parameters, and automatic reports should be simplified and improved for easier use and reduced opportunity for user error.

References

[1] "Parkinson's disease: Hope through research." National Institute of Neurological Disorders and Stroke, NIH Publication No. 06-139, January 2006.

[2] K. Noyes, H. Liu, Y. Li, R. Holloway, and A. W. Dick, "Economic burden associated with Parkinson's disease on elderly medicare beneficiaries," *Movement Disorders*, vol. 21, pp. 362–372, Mar 2006.

[3] D. M. Huse, K. Schulman, L. Orsini, J. Castelli-Haley, S. Kennedy, and G. Lenhart, "Burden of illness in Parkinson's disease," *Movement Disorders*, vol. 20, pp. 1449–1454, Nov 2005.

[4] Goetz and Stebbins, "Assuring interrater reliability for the UPDRS motor section: Utility of the UPDRS teaching tape," *Movement Disorders*, vol. 19, no. 12, pp. 1453–1456, 2004.

[5] R. Pahwa, K. E. Lyons, and W. C. Koller, *Handbook of Parkinson's Disease*. Marcel Dekker, Inc., 2003.

[6] J. C. Hobart, D. L. Lamping, and A. J. Thompson, "Evaluating neurological outcome measures: the bare essentials," *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 60, pp. 127–130, Feb 1996.

[7] J. McNames, "Rapid development and validation of objective measures of Parkinson's disease." Memo to Kinetics Foundation, Los Altos, California, March 2009.

[8] Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease, "The Unified Parkinson's Disease Rating Scale (UPDRS): status and recommendations," *Movement Disorders*, vol. 18, pp. 738–750, Jul 2003.

[9] O. Rascol, C. Goetz, W. Koller, W. Poewe, and C. Sampaio, "Treatment interventions for Parkinsons disease: an evidence based assessment," *THE LANCET*, vol. 359, pp. 1589–1598, May 2002.

[10] H. M. Bronte-Stewart, L. Ding, C. Alexander, Y. Zhou, and G. P. Moore, "Quantitative digitography (qdg): a sensitive measure of digital motor control in idiopathic parkinson's disease.," *Mov Disord*, vol. 15, pp. 36–47, Jan 2000.

[11] A. L. T. Tavares, G. S. X. E. Jefferis, M. Koop, B. C. Hill, T. Hastie, G. Heit, and H. M. Bronte-Stewart, "Quantitative measurements of alternating finger tapping in parkinson's disease correlate with updrs motor disability and reveal the improvement in fine motor control from medication and deep brain stimulation.," *Mov Disord*, vol. 20, pp. 1286–1298, Oct 2005.

[12] M. M. Koop, N. Shivitz, and H. Bronte-Stewart, "Quantitative measures of fine motor, limb, and postural bradykinesia in very early stage, untreated parkinson's disease.," *Mov Disord*, vol. 23, pp. 1262–1268, Jul 2008.

[13] C. G. Goetz, G. T. Stebbins, D. Wolff, W. DeLeeuw, H. Bronte-Stewart, R. Elble, M. Hallett, J. Nutt, L. Ramig, T. Sanger, A. D. Wu, P. H. Kraus, L. M. Blasucci, E. A. Shamim, K. D. Sethi, J. Spielman, K. Kubota, A. S. Grove, E. Dishman, and C. B. Taylor, "Testing objective measures of motor impairment in early Parkinson's disease: Feasibility study of an at-home testing device," *Movement Disorders*, vol. 24, pp. 551–556, Mar 2009.

[14] L. M. Harvill, "An ncme instructional module on standard error of measurement," *Educational Measurement: Issues and Practice*, vol. 10, no. 2, pp. 181–189, 1991.

[15] J. M. Bland and D. G. Altman, "Statistics notes: Cronbach's alpha," *British Medical Journal*, vol. 314, p. 572, February 1997.

[16] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients," *Psychological Methods*, vol. 1, pp. 30–46, Mar. 1996. Corrections in no. 4, pg. 390.

[17] L. V. Metman, B. Myre, N. Verwey, S. Hassin-Baer, J. Arzbaecher, D. Sierens, and R. Bakay, "Test-retest reliability of updrs-iii, dyskinesia scales, and timed motor tests in patients with advanced Parkinsons disease: An argument against multiple baseline assessments," *Movement Disorders*, vol. 19, no. 9, pp. 1079–1084, 2004.

[18] R. L. Rosnow and R. Rosenthal, "Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers," *Psychological Methods*, vol. 1, no. 4, pp. 331–340, 1996.

[19] J. Parkinson, "An essay on the shaking palsy," *The Journal of Neuropsychiatry and Clinical Neurosciences*, vol. 14, no. 2, pp. 223–236, 2002.

[20] S. Fahn, R. L. Elton, and UPDRS Development Committee, "Unified Parkinsons disease rating scale," in *Recent developments in Parkinsons disease* (S. Fahn, C. D. Marsden, M. Goldstein, and D. B. Calne, eds.), vol. 2, pp. 153–163, 293–304, Florham Park, N. J.: Macmillan Healthcare Information, 1987.

[21] C. G. Goetz, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, G. T. Stebbins, M. B. Stern, B. C. Tilley, R. Dodel, B. Dubois, R. Holloway, J. Jankovic, J. Kulisevsky, A. E. Lang, A. Lees, S. Leurgans, P. A. LeWitt, D. Nyenhuis, C. W. Olanow, O. Rascol, A. Schrag, J. A. Teresi, J. J. V. Hilten, and N. LaPelle, "Movement disorder society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Process, format, and clinimetric testing plan," *Movement Disorders*, vol. 22, pp. 41–47, Jan 2007.

[22] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel, B. Dubois, R. Holloway, J. Jankovic, J. Kulisevsky, A. E. Lang, A. Lees, S. Leurgans, P. A. LeWitt, D. Nyenhuis, C. W. Olanow, O. Rascol, A. Schrag, J. A. Teresi, J. J. van Hilten, N. LaPelle, and Movement Disorder Society UPDRS Revision Task Force, "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results," *Mov Disord*, vol. 23, pp. 2129–2170, Nov 2008.

[23] P. Martinez-Martin, A. Gil-Nagel, L. M. Gracia, J. B. Gomez, J. Martinez-Sarries, F. Bermejo, and T. C. M. Group, "Unified Parkinsons disease rating scale UPDRS characteristics and structure," *Movement Disorders*, vol. 9, no. 1, pp. 76–83, 1994.

[24] G. T. Stebbins and C. G. Goetz, "Factor structure of the unified Parkinson's disease rating scale: Motor examination section.," *Movement Disorders*, vol. 13, pp. 633–636, Jul 1998.

[25] G. T. Stebbins, C. G. Goetz, A. E. Lang, and E. Cubo, "Factor analysis of the motor section of the unified Parkinson's disease rating scale during the off-state.," *Movement Disorders*, vol. 14, pp. 585–589, Jul 1999.

[26] J. M. Rabey, H. Bass, U. Bonuccelli, D. Brooks, and P. Klotz, "Evaluation of the short Parkinson's evaluation scale: A new friendly scale for the evaluation of Parkinson's disease in clinical drug trials," *Clinical Neuropharmacology*, vol. 20, no. 4, pp. 322–337, 1997.

[27] M. Richards, K. Marder, L. Cote, and R. Mayeux, "Interrater reliability of the unified Parkinson's disease rating scale motor examination," *Movement Disorders*, vol. 9, no. 1, pp. 89–91, 1994.

[28] R. Camicioli, S. J. Grossmann, P. S. Spencer, K. Hudnell, and W. K. Anger, "Discriminating mild parkinsonism: Methods for epidemiological research," *Movement Disorders*, vol. 16, pp. 33–40, 2001.

[29] D. Bennett, K. Shannon, L. Beckett, C. Goetz, and R. Wilson, "Metric properties of nurses' ratings of parkinsonian signs with a modified unified Parkinson's disease rating scale," *Neurology*, vol. 49, pp. 1580–7, December 1997.

[30] A. Siderowf, M. McDermott, K. Kieburtz, K. Blindauer, S. Plumb, I. Shoulson, and the Parkinson Study Group, "Test-retest reliability of the unified Parkinsons disease rating scale UPDRS in patients with early Parkinsons disease: Results from a multicenter clinical trial," *Movement Disorders*, vol. 17, no. 4, pp. 758–763, 2002.

[31] T. Steffen and M. Seney, "Test-retest reliability and minimal detectable change on balance and ambulation tests, the 36-item short-form health survey, and the unified Parkinson disease rating scale in people with parkinsonism," *Physical Therapy*, vol. 88, pp. 733–746, Jun 2008.

[32] P. Martinez-Martin and M. J. Forjaz, "Metric attributes of the unified Parkinson's disease rating scale 3.0 battery: Part i, feasibility, scaling assumptions, reliability, and precision.," *Movement Disorders*, vol. 21, pp. 1182–1188, Aug 2006.

[33] C. Goetz, W. Poewe, O. Rascol, and C. Sampaio, "Research review: Evidence-based medical review update: Pharmacological and surgical treatments of Parkinsons disease: 2001 to 2004," *Movement Disorders*, vol. 20, no. 5, pp. 523–539, 2005.

[34] R. Hauser, P. Auinger, and D. Oakes, "Levodopa response in early Parkinsons disease," *Movement Disorders*, vol. 24, no. 16, p. 23282336, 2009.

[35] L. Buddenberg and C. Davis, "Test-retest relaibility of the Purdue pegboard test," *American Journal of Occupational Therapy*, vol. 54, pp. 555–558, September/October 2000.

[36] J. R. Reddon, D. M. Gill, S. E. Gauk, and M. D. Maerz, "Purdue pegboard: Test-retest estimates," *Perceptual and Motor Skills*, vol. 66, pp. 503–506, 1988.

[37] A. Samii, I. M. Turnbull, A. Kishore, M. Schulzer, E. Mak, S. Yardley, and D. B. Calne, "Reassessment of unilateral pallidotomy in Parkinsons disease a 2-year follow-up study," *Brain*, vol. 122, pp. 417–425, 1999.

[38] C. A. Haaxma, B. R. Bloem, G. F. Borm, and M. W. I. M. Horstink, "Comparison of a timed motor test battery to the unified Parkinson's disease rating scale-iii in Parkinson's disease.," *Movement Disorders*, vol. 23, pp. 1707–1717, Sep 2008.

[39] P. H. Kraus, P. Klotz, A. Hoffmann, J. Lewe, and H. Przuntek, "Analysis of the course of Parkinsons disease under dopaminergic therapy: Performance of fast tapping is not a suitable parameter," *Movement Disorders*, vol. 20, no. 3, 2005.

[40] J. G. Nutt, E. S. Lea, L. V. Houten, R. A. Schuff, and G. J. Sexton, "Determinants of tapping speed in normal control subjects and subjects with Parkinson's disease: Differing effects of brief and continued practice," *Movement Disorders*, vol. 15, no. 5, pp. 843–849, 2000.

[41] J. Ghika, A. W. Wiegner, J. J. Fang, L. Davies, R. R. Young, and J. H. Growdon, "Portable system for quantifying motor abnormalities in Parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 40, pp. 276–283, Mar 1993.

[42] R. J. Dunnewold, C. E. Jacobi, and J. J. van Hilten, "Quantitative assessment of bradykinesia in patients with Parkinson's disease," *Journal of Neuroscience Methods*, vol. 74, pp. 107–112, Jun 1997.

[43] S. Louie, M. M. Koop, A. Frenklach, and H. Bronte-Stewart, "Quantitative lateralized measures of bradykinesia at different stages of Parkinsons disease: The role of the less affected side," *Movement Disorders*, vol. 24, no. 13, 2009.

[44] M. M. Koop, A. Andrzejewski, B. C. Hill, G. Heit, and H. M. Bronte-Stewart, "Improvement in a quantitative measure of bradykinesia after microelectrode recording in patients with Parkinsons disease during deep brain stimulation surgery," *Movement Disorders*, vol. 21, no. 5, pp. 673–678, 2006.

[45] G. Giovannoni, J. van Schalkwyk, V. U. Fritz, and A. J. Lees, "Bradykinesia akinesia inco-ordination test (BRAIN TEST): an objective computerised assessment of upper limb motor function," *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 67, pp. 624–629, Nov 1999.

[46] J. Jankovic, L. Ben-Arie, K. Schwartz, K. Chen, M. Khan, E. C. Lai, J. K. Krauss, and R. Grossman, "Movement and reaction times and fine coordination tasks following pallidotomy," *Movement Disorders*, vol. 14, no. 1, pp. 57–62, 1999.

[47] Burns and DeJong, "A preliminary report on the measurement of Parkinsons disease.," *Neurology*, vol. 10, p. 1096, 1960.

[48] J. Cecil Hastings, F. Mosteller, J. W. Tukey, and C. P. Winsor, "Low moments for small samples: A comparative study of order statistics," *The Annals of Mathematical Statistics*, vol. 18, pp. 413–426, September 1947.

[49] B. Efron, "Bootstrap methods: Another look at the jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.

[50] E. L. Stegemöller, T. Simuni, and C. MacKinnon, "Effect of movement frequency on repetitive finger movements in patients with Parkinson's disease," *Movement Disorders*, vol. 24, pp. 1162–1169, Jun 2009.