Detection of Motor Impairment in Parkinson's Disease Via Mobile Touchscreen Typing

Teresa Arroyo-Gallego, María Jesus Ledesma-Carbayo, Álvaro Sánchez-Ferro, Ian Butterworth, Carlos S. Mendoza, Michele Matarazzo, Paloma Montero, Roberto López-Blanco, Verónica Puertas-Martín, Rocío Trincado, and Luca Giancardo*

***Abstract*—Mobile technology is opening a wide range of opportunities for transforming the standard of care for chronic disorders. Using smartphones as tools for longitudinally tracking symptoms could enable personalization of drug regimens and improve patient monitoring. Parkinson's disease (PD) is an ideal candidate for these tools. At present, evaluation of PD signs requires trained experts to quantify motor impairment in the clinic, limiting the frequency and quality of the information available for understanding the status and progression of the disease. Mobile technology can help clinical decision making by completing the information of motor status between hospital visits. This paper presents an algorithm to detect PD by analyzing the typing activity on smartphones independently of the content of the typed text. We propose a set of touchscreen typing features based on a covariance, skewness, and kurtosis analysis of the timing information of the data to capture PD motor signs. We tested these features, both independently and in a multivariate framework, in a population of 21 PD and 23 control subjects, achieving a sensitivity/specificity of $0.81/0.81$ for the best performing feature and $0.73/0.84$ for the best multivariate method. The results of the alternating finger-tapping, an established motor test, measured in our cohort are $0.75/0.78$. This paper contributes to the development of a home-based, high-compliance, and high-frequency PD motor test by analysis of routine typing on touchscreens.**

***Index Terms*—Feature extraction, finger tapping, keystroke dynamics, mHealth, passive monitoring, signal processing, smartphone.**

T. Arroyo-Gallego is with the Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA, with the Biomedical Image Technologies, Universidad Politécnica de Madrid, and also with CIBER-BBN, Madrid, Spain.

M. J. Ledesma-Carbayo is with the the Biomedical Image Technologies, Universidad Politécnica de Madrid, and also with CIBER-BBN, Madrid, Spain.

A. Sánchez-Ferro is with the Madrid-MIT M+Visión Consortium, Research Laboratory of Electronics, Massachusetts Institute of Technology, and also with HM Hospitales—Centro Integral en Neurociencias HM CINAC, Spain.

I. Butterworth is with Madrid-MIT M+Visión Consortium, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA.

C. S. Mendoza is with Asana Weartech, Spain and also with Madrid-MIT M+Visión Consortium, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA.

M. Matarazzo is with the HM Hospitales—Centro Integral en Neurociencias HM CINAC, and also with the Instituto de Investigación Hospital 12 de Octubre (i+12), Madrid, Spain.

P. Montero is with the Movement Disorders Unit, Hospital Clinico San Carlos, Madrid, Spain.

R. López-Blanco, V. Puertas-Martín, and R. Trincado are with the Instituto de Investigación Hospital 12 de Octubre (i+12), Madrid, Spain.

*L. Giancardo is with the Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030 USA, and also with the Madrid-MIT M+Visión Consortium, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: luca.giancardo@uth.tmc.edu).

I. Introduction

PARKINSON'S disease (PD) is a chronic neurological disorder causing progressive disability related to the loss of nigrostriatal dopaminergic neurons. It is the second most common neurodegenerative disorder, presenting an annual incidence rate of 8-18 per 100,000 persons [1]. PD is commonly defined by motor impairment, involving tremor, bradykinesia, postural instability, gait difficulty, and rigidity. However, non-motor signs, such as mood alteration, cognitive alteration or sleep disturbances, are also characteristic of this disease [2]. Symptom diversity affects patients' daily life in all physical, social and mental planes, producing an adverse impact in the main components of health-related quality of life [3].

At this time, available medication provides symptomatic relief by setting an appropriate balance of dopamine levels. One of the main difficulties in adapting treatment parameters is the lack of a clear and objective measurement method to accurately quantify and monitor the disease stage for each individual case [4]. The Unified Parkinson's Disease Rating Scale (UPDRS) is the most commonly used metric in the clinical evaluation of PD. It consists of a standardized test that provides an overall score of the patients' functional capabilities [5]. The UPDRS-III evaluates motor performance by having the subject undertake a series of motor tasks and assigning each a score from 0 to 4. The total UPDRS-III score is calculated as the sum of all the individual task scores. Despite UPDRS-III being the most accepted standard in clinical assessment, it requires significant training to minimize inter-rater variability [6]. The need for an experienced clinician to subjectively evaluate the progress of the disease typically limits the gathering of information to

on-site medical examinations. This clinical data constitutes the main basis on which clinicians adjust patients therapy, which means that decision-making is subject to the participant's recall-bias and is based on limited information. In summary, current practices not only lack better monitoring of PD progress, but also do not provide a consistent and objective evaluation of the measured signs.

Finger-tapping tests [7] are complementary methods that provide additional and objective information about motor function health. These tests employ standardized finger-movement exercises to detect and quantify psychomotor dysfunction. Alternating finger-tapping (AFT) is one of the varieties of this method. Using a single hand, the tested subject has to alternatively press two specified buttons as fast as possible during a predefined time [8]. The test is repeated for both hands and the final score is the average number of pressed keys between the two. Despite its simplicity, AFT is a commonly used method to evaluate PD as it provides useful information to characterize upper limb motor function [9].

An increasing interest in developing new ways to apply technology for creating objective clinical assessment tools is shared by patients, clinicians, and researchers. In the particular case of PD, recent reviews confirm a keen interest in the exploration of technological improvements in patient care [10]. In [11], authors present a survey highlighting the existing consensus between clinicians and patients about the need for a monitoring system to better understand response to therapy and improve treatment titration.

PD motor impairment manifests in a variety of ways, which allows for a broad range of measurement methods. Consequently, a variety of techniques to provide complementary information for optimizing PD care are emerging [12]. Reported results on ambulatory monitoring of PD patient activity shows great promise [13]. However, its translation to clinical practice remains elusive. The application of accelerometers and other sensing systems to develop high frequency motor tracking tools has become one of the main trends thanks to the advances in sensor miniaturization, wireless technology, signal processing and data analysis [14]–[16]. A notable challenge in sensor-based solutions is the development of advanced algorithms to evaluate the highly complex patterns that result from the interference of PD motor signs and normal daily activity movements, this calls for advancement in algorithms to process the accelerometers data [17]. In the recent years, the use of commodity hardware such as smartphones has gained traction over systems using specialized sensors. For instance, the microphone of these devices has been used to predict PD severity via speech analysis algorithms [18].

Touchscreens and embedded accelerometers are another source of data to quantify PD signs. One of the largest studies to date is mPower [19]. A smartphone-based activity tracker including touchscreen tapping, memory, voice, posture and gait tests that collected longitudinal data from a large number of PD patients and controls during a six month period.

A common limitation between the cited tools is that they require subjects' active participation, in the sense that subjects need to be reminded to take each test. This leads to reduced compliance. In the mPower initiative, Bot *et al.* [19] reported that out of the 9,520 participants who opted to share broadly their data, less than 10% (898) performed the finger tapping test for 5 days or more.

We propose a solution that takes advantage of the ubiquity and pervasiveness of smartphone technology. Importantly, in contrast to many other mobile-based approaches, our solution is transparent to the user, and does not require the user to take any action to initiate a test. Our primary objective is for this transparent monitoring to provide information comparable to current motor tests. More specifically our approach should simplify the monitoring process by passively collecting information from the routine use of smartphone devices. In our previous work [20], we have demonstrated that clinically relevant motor function changes can be measured by timing key press/release events during typing on physical keyboards, irrespective of language or text typed. In the specific case of PD, we have shown that daily interaction with physical keyboards can be used to measure motor signs in the early stages of the disease [21]. In this work, we introduce a set of numerical features derived from similar keystroke dynamic variables on mobile phone touchscreens. We learn characteristic PD typing patterns to facilitate detection and quantification of the motor signs related to this disease. PD motor phenotype is described by slowness, lack of spontaneous movement, rigidity, and tremor. This clinical picture should affect the unconstrained finger performance while interacting with smartphone devices.

In this paper, we propose a smartphone-based approach to assess PD motor signs. The developed solution uses touchscreens as hardware support, and relies on the typing signal as input to evaluate motor function anomalies. Our study is a first step towards a transparent and ubiquitous motor sign assessment method that is objective, convenient, and can produce quasi-continuous ambulatory data. The main contribution of this paper is a new methodology to quantify motor impairment through the analysis of the typing signal collected via smartphone devices. We tested our solution on a validation cohort that includes data from 21 PD and 23 control subjects. The performance and relevance of the developed tool is verified by comparing the obtained results with respect to the alternating finger-tapping (AFT) motor test.

II. Materials and Methods

This sections includes a general description of the data acquisition, followed by the presentation of the proposed methodology.

A. Data Acquisition

We collected 51 typing signals from a population composed of 24 people diagnosed with Parkinson's and 27 healthy controls. Subjects gave informed consent prior to experiments, and experimental procedures were approved by the Committee On the Use of Humans as Experimental Subjects (COUHES) at the Massachusetts Institute of Technology, protocol no. 1504007090. During the meeting the subjects independently underwent a clinical evaluation including the UPDRS-III test

TABLE I
DATASET DEMOGRAPHICS

	Avg. (std) Parkinson's	Avg. (std) Controls	Significance
Age	59.24 (11.43)	54.35 (13.95)	$p = 0.32$
Women # (%)	11 (52%)	19 (83%)	$p < 0.05$
Men # (%)	10 (48%)	4 (17%)	$p < 0.05$
UPDRS-III	17.76 (7.92)	1.22 (1.70)	$p < 0.001$
Alternating finger-tapping	49.17 (10.65)	67.54 (14.11)	$p < 0.001$
Hoehn and Yahr	2.05 (0.31)	N.A.	N.A.
n (total n = 44)	21	23	

The complete study cohort comprised 51 subjects. From the total participants, 44 provided enough typing information to perform the analysis. A minimum of 5 key presses every 15 seconds during at least half of the duration of the typing task was required to apply the proposed method. Seven subjects, 3 from the Parkinson's group and 4 healthy controls, did not provide enough data and were excluded from the analysis (see Materials and Methods). The table provides a summary of the demographic information of the participants included in the analysis, 21 people diagnosed with Parkinson's (PD) and 23 control subjects (CNT). PD subjects and controls are statistically similar in age, according to the two-sided Mann-Whitney U test. The same test suggests gender might be a confound variable in this study. These differences were accounted in the analysis. The table also shows the results of the clinical evaluation that includes UPDRS-III, alternating finger-tapping and the Modified Hoehn and Yahr scale. The Hoehn and Yahr scale is a widely used clinical rating scale that defines broad categories of disability in PD in a 0 to 5 range.

conducted by a movement disorder expert. After the clinical assessment, each participant created an account on our website (www.neuroqwerty.com). The Alternating finger-tapping (AFT) test was performed on a physical keyboard. Subjects had to alternatively press two keys, separated approximately 25 cm, using their index finger. They repeated the test for both hands. The final score was computed as the average number of buttons pressed between the two hands. The typing data was collected using dedicated smartphone software. Participants transcribed a randomly-selected text excerpt for five minutes and were instructed to type as they would normally do in order to reflect actual routine use of the device. PD subjects were tested during their "ON" state, under best medical treatment.

Seven participants, 4 from the control group and 3 Parkinson's subjects, did not have enough data to compute the feature analysis and were excluded from the dataset. All of them presented a typing rate below 20 keys per minute for at least half of the typing time. Table I summarizes the demographic information of the remaining 44 subjects that were included in the analysis.

For the study, we developed a custom screen keyboard in order to enable typing data collection. The application was based on the open source software keyboard AnySoftKeyboard (github.com/AnySoftKeyboard). Running in the background of any application that receives keyboard input, it captures the time stamps corresponding to press and release events for each keystroke. The system tested has a clock speed and a theoretical low-level sampling frequency of 1.2 GHz. Our implementation uses a software timer with a time granularity of 1 millisecond. The encrypted information was sent to a remote server for the analysis. All the subjects were tested on an Android terminal, i.e. Motorola Moto G II running Android 5.0. In Fig. 1 we provide a graphical representation of the study procedure.

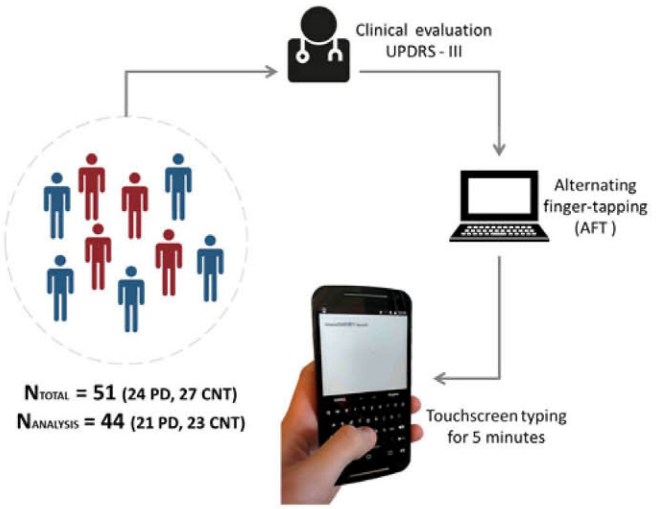

Fig. 1. The figure presents a schema of the study procedure that comprises a clinical evaluation, finger-tapping test and our typing test. For the clinical evaluation a movement disorder specialist filled-in the motor section in the Unified Parkinsons Rating Scale (UPDRS-III). The alternating finger-tapping test was included as an external reference to quantify upper limbs dexterity. It was performed on a physical keyboard. The typing test consisted of a five minutes task were participants were asked to transcribe a non-standardized text excerpt using a touchscreen device. The custom screen keyboard and smartphone model used in the test are shown in this figure.

B. Data Analysis

The method description is divided into three different phases as follows: An initial signal conditioning phase in order to minimize signal noise and artifacts. Then, statistical analysis is used to describe the processed signal using a limited number of typing features. In the last stage the feature vector is evaluated to determine its suitability for detecting PD status.

1) Signal Conditioning: In this study, we define the typing signal ($X[t]$) as the sequence of flight time (FT) values corresponding to each key tap. In the context of our work, we define FT as the release latency between key taps, i.e. for two consecutive keystrokes the time measured between first and second key release times. The captured typing data requires further processing in order to remove noise, minimize the effect of confounding factors on the analysis and define a standardized representation of the whole signal. Noise can be introduced by many sources, such as software inaccuracies or unnatural typing episodes (e.g. special keys). Additionally, to reduce noise levels, the signal is processed by a series of conditional filters that remove potentially noisy samples if the FT value exceeds a 3-second threshold, or if they correspond to special key-types that engage non-standard digit kinematics (e.g. SHIFT). To minimize the effect of typing skills in the results of the analysis, each signal is normalized by subtracting its mean value to every data sample in $X[t]$. Fig. 2 compares the probability density functions for the normalized FT (NFT) data grouped by condition.

$$X'[t] = X[t] - \bar{X} : X' \in [\theta_A, \theta_B]$$

where X' is the normalized signal and $\bar{X}$ its average. The value t represents the time at which a key has been pressed to generate

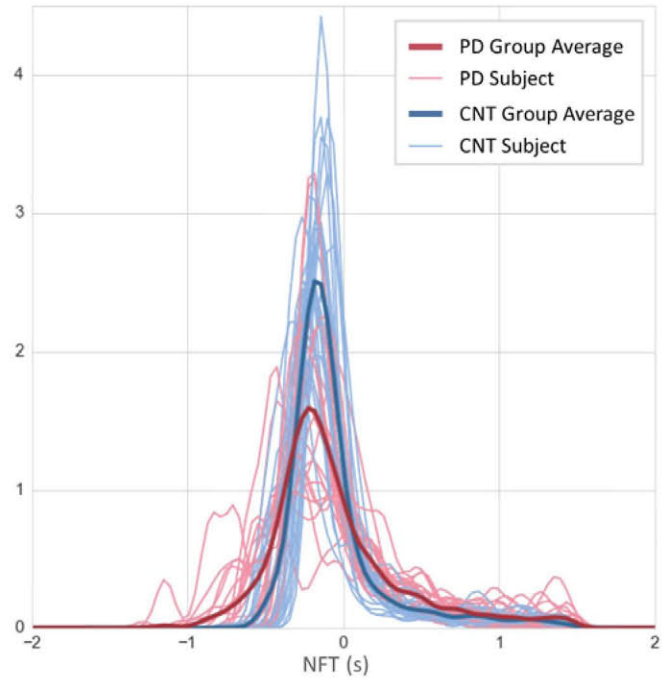

Fig. 2. The figure shows the probability density distribution (PDF) estimated for the normalized flight time (NFT) signals both for each subject (light color) and grouped by condition (dark color). Normalization minimizes the influence of confounding variables related to typing skills. Parkinson's subjects' (PD) distributions present a greater sparsity than controls' (CNT) distributions. A Mann-Whitney U test suggests a significant difference between the NFT values measured on PD participants and controls ($p < 0.001$).

the relative FT signal. The parameters θ_A and θ_B give the estimated range of values in which the 99% of the normalized FT data is concentrated. These two parameters have been estimated in an external typing database of 27 healthy subjects.

We define a new type of signal representation (X'_S) based on the FT time series as to adapt the normalized data to the following analysis stages. Given $X'[t]$, the signal structure, X'_S, is defined as a set of vectors X'_{Si} with a varying number of elements but a fixed length in the time domain:

$$X'_{Si}[t, N] = X'[t]w[t - iN]$$

where i is a strictly positive integer which serves as an index to the list of vectors, $N = 15,000$ is the length of the window time expressed in milliseconds and $w[n]$ is defined as:

$$w[n] = \begin{cases} 1, & 0 \leq n < N \\ 0, & \text{otherwise} \end{cases}$$

2) Feature Extraction: Evaluating X'_S using distribution and covariance based approaches we define two different feature families.

a) Skewness and Kurtosis: These measurements correspond to the third and forth moments of a distribution. Skewness can be interpreted as an indicator of distribution symmetry, while kurtosis measures the variable distribution flatness. Each element in the typing structure (X'_{Si}) is evaluated as an independent realization of the same random variable, with its corresponding distribution that we will call sub-distribution in the context of $X[t]$. Then, for each sub-distribution a pair including skewness (Sk_i) and kurtosis (Kt_i) descriptors are computed.

For a sample of n values, a natural method of moments estimator of the population skewness is:

$$Sk_i = \frac{\sum_{m=1}^{M_i} \left(X'_{Si}[m] - \bar{X}'_{Si}\right)^3}{\sigma^3_{Si}}$$

For a sample of n values the sample excess kurtosis is:

$$Kt_i = \frac{\sum_{m=1}^{M_i} \left(X'_{Si}[m] - \bar{X}'_{Si}\right)^4}{\sigma^4_{Si}} - 3$$

where M_i is the length of the i^{th} vector in X'_S and σ_{Si} is the standard deviation of X'_{Si}.

With I being the number of sub-distributions that compose the structured typing signal, the analysis described above generates a total of I measures for each metric. These values are reduced to four final features computed as the average and standard deviation of the I skewness measurements $(\bar{Sk}, \sigma_{Sk})$ and the I kurtosis measurements $(\bar{Kt}, \sigma_{Kt})$.

b) Covariance: The typing signal structure (X'_S) is transformed into a matrix (H) by applying Kernel Density Estimation (KDE). A similar approach was presented in [20] to define the Key Hold Time Evolution Matrix.

We apply KDE to estimate the probability density function (PDF) that represents the underlying distribution of each element in the typing structure. Given a typing sub-sample X'_{Si} of size M_i, its PDF f_i is computed as follows:

$$f_i(y, b) = \sum_{m=1}^{M_i} K((y - X'_{Si}[m])/b)$$

where b is a bandwidth parameter that controls K, a Gaussian kernel:

$$K(x, b) \propto exp\left(-\frac{x^2}{2b^2}\right)$$

Each function f_i is quantizied using pre-defined mapping levels $\vec{v}$. This allows a standardized $I \times L$, matrix representation of the typing signal as:

$$H_{i,j} = f_i(\vec{v}[j])$$

We used our external dataset, not used for training or testing, comprised by 27 healthy subjects, to adjust the value of the bandwidth parameter b [22], as well as the number of quantization levels $(L = 10)$ and the limits of the mapping vector $(\vec{v})$.

The corresponding covariance matrix (COV_H) is estimated for the resulting NFT distribution matrix (H) as follows:

$$COV_{H_{i,j}} = cov(H_{i*}, H_{j*}) =$$

$$= \frac{1}{L-1} \sum_{l=1}^{L} \left(H_{i,l} - \bar{H}_{i*}\right)\left(H_{j,l} - \bar{H}_{j*}\right)$$

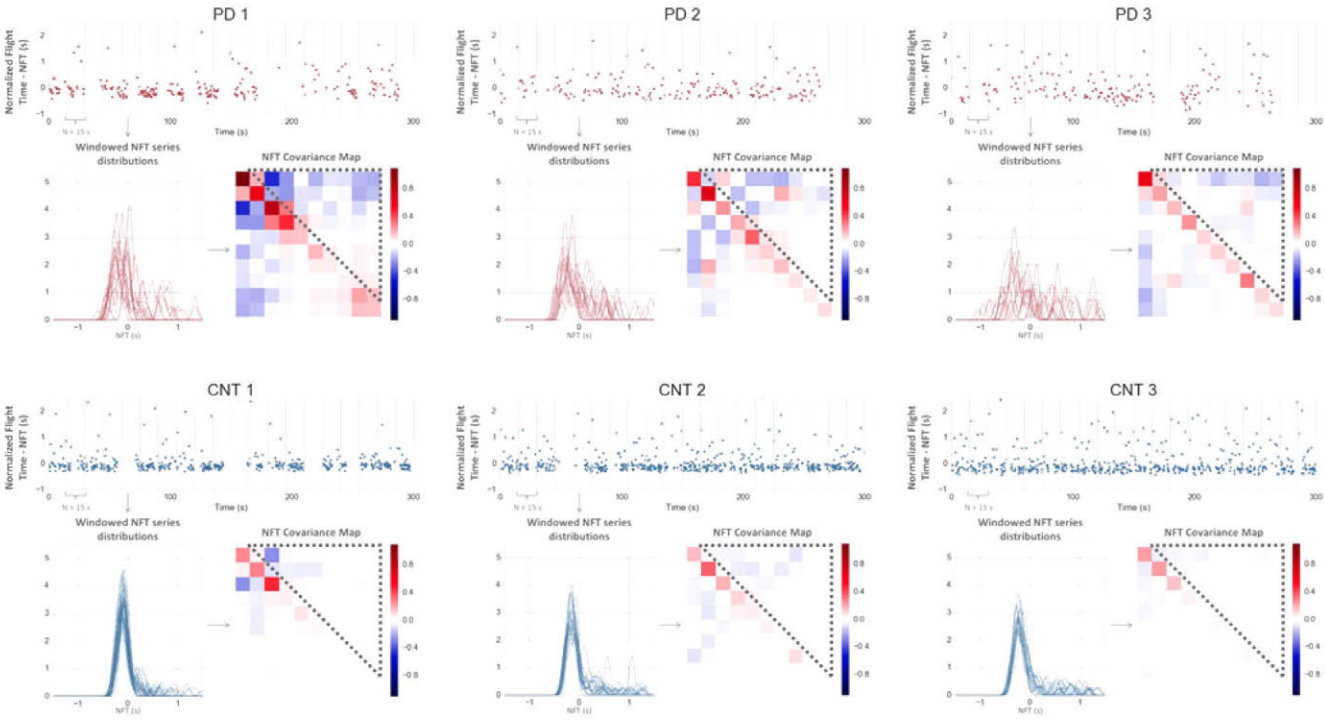

Fig. 3. The figure compares six signal examples from three people diagnosed with Parkinson's (PD) and three healthy controls (CNT). The normalized flight time series (NFT) is split into 15 second-length windows. Applying Kernel Density Estimation (KDE) we compute the sub-distribution representing the information contained in each window. The mean and standard deviation of the skewness and kurtosis values measured on each sub-distribution define four of the features ($\bar{S}k$, σ_{Sk}, $\bar{K}t$, σ_{Kt}) that are included in the final 7-dimensional feature vector. The NFT covariance map represents the correlation across the NFT sub-distributions. We define the covariance vector (Cv) as an array including the coefficients in the strict upper triangle of the covariance matrix, i.e. above the matrix main diagonal. We extract three metrics from the covariance analysis that complete the typing feature vector ($\bar{C}v$, σ_{Cv}, $\sum|Cv|$) Distributions show a higher uniformity of the NFT values for the CNT's signals compared to PD's. Covariance maps for PD show stronger correlation and anti-correlation within sub-distributions while CNT's maps present values nearer 0 for the entire matrix.

TABLE II
RESULTS, UNIVARIATE ANALYSIS (INDEPENDENT TYPING FEATURES)

Feature	Avg. (std) Parkinson's	Avg. (std) Controls	AUC [5%, 95%]	Significance
$\bar{S}k$	0.955 (0.572)	1.742 (0.536)	0.85 [0.74, 0.95]	$p < 0.001$
σ_{Sk}	0.652 (0.135)	0.837 (0.210)	0.77 [0.64, 0.87]	$p < 0.01$
$\bar{K}t$	0.767 (1.749)	3.587 (2.594)	0.87 [0.78, 0.95]	$p < 0.001$
σ_{Kt}	1.691 (0.977)	3.595 (1.593)	0.88 [0.78, 0.95]	$p < 0.001$
$\bar{C}v$	−0.019 (0.015)	−0.018 (0.006)	0.46 [0.38, 0.69]	$p = 0.66$
σ_{Cv}	0.104 (0.030)	0.068 (0.029)	0.84 [0.72, 0.93]	$p < 0.001$
$\sum\|Cv\|$	3.290 (1.071)	1.839 (0.656)	0.91 [0.82, 0.97]	$p < 0.001$
n (total n = 44)	21	23		

The table shows the mean values and performance of the typing features and the reference metrics, including the ROC AUC mean and confidence intervals achieved by each measurement and the results of the Mann-Whitney U test to analyze if the null hypothesis, that Parkinson's disease (PD) and control (CNT) subjects come from the same population, can be rejected. Covariance sum $\sum|Cv|$ presents the best discrimination performance with an AUC of 0.91 and significance $p < 0.001$.

where H_{i*} is a vector that contains the H matrix values for the i^{th} row:

$$H_{i*} = [H_{i,1}, H_{i,2}, \ldots, H_{i,l}]^T$$

Being Cv a covariance vector including the upper triangle elements of COV_H, i.e. the coefficients in the upper portion above the main diagonal of the matrix, we define the covariance typing features as follows: covariance mean ($\bar{C}v$), covariance standard deviation (σ_{Cv}) and the sum of the absolute values of the covariance vector elements ($\sum|Cv|$). A graphic representation of the typing signal characterization is presented in Fig. 3.

3) Evaluation Methodology

The proposed features are based on a limited set of parameters that are estimated on an external dataset of 27 healthy subjects. These parameters, shown in the previous sections, provide a general description of the typing signal and are independent of motor function status, i.e. they are not optimized to enhance the separation between Parkinson's participants and controls.

First, we assess the classification performance of the proposed typing features with univariate methods. Next, we evaluate the joint discriminatory power of the features in a multivariate analysis framework.

The multivariate method assembles a feature selection transform followed by a final estimator. We use a nested-cross validation strategy for performance evaluation, i.e. a combination of two embedded cross-validation loops. The inner k-fold cross-validation loop is used to identify the relevant features and estimate the model hyperparameters based on the training folds of the outer leave-one-out cross-validation fold. The outer loop

TABLE III
RESULTS, MULTIVARIATE ANALYSIS (AGGREGATED TYPING FEATURES)

Model	Feature Selection	Classifier	Avg. Score (std) Parkinson's	Avg. Score (std) Controls	AUC [5%,95%]	Significance
1	L1 (Lasso)	Logistic Regression	2.127 (2.883)	−1.684 (2.127)	0.87 [0.75, 0.94]	$p < 0.001$
2	L1 (Lasso)	Linear SVM	0.825 (1.109)	−0.663 (0.730)	0.88 [0.78, 0.95]	$p < 0.001$
3	L1 (Lasso)	AdaBoost	3.603 (6.612)	−0.218 (1.591)	0.73 [0.58, 0.84]	$p < 0.01$
4	L2 (Ridge)	Logistic Regression	1.767 (2.870)	−2.241 (2.293)	0.87 [0.77, 0.95]	$p < 0.001$
5	L2 (Ridge)	Linear SVM	0.736 (1.063)	−0.709 (0.926)	0.86 [0.74, 0.95]	$p < 0.001$
6	L2 (Ridge)	AdaBoost	0.617 (1.599)	−1.464 (1.734)	0.82 [0.69, 0.92]	$p < 0.001$
7	Gini Impurity	Logistic Regression	1.434 (2.350)	−1.236 (1.850)	0.81 [0.67, 0.91]	$p < 0.001$
8	Gini Impurity	Linear SVM	0.650 (1.035)	−0.588 (0.683)	0.86 [0.74, 0.95]	$p < 0.001$
9	Gini Impurity	AdaBoost	2.131 (5.123)	−1.019 (1.972)	0.80 [0.66, 0.90]	$p < 0.05$
n (total n = 44)	21	23				

The table summarizes the results of the multivariate analysis. We evaluate the classification performance of different models that aggregate the information of the proposed typing features. We tested a total of nine models, built as the possible combinations of three different feature selection methods and three estimators. A nested cross validation framework was implemented to train and test the models. For each model, we include the mean and confidence intervals of the AUC and the results of the Mann-Whitney U test to reject the null hypothesis that Parkinson's disease (PD) and control (CNT) subjects come from the same population. Model 2, a linear support vector classifier preceded by L1-regularized linear model (Lasso) for feature selection, presents the best discrimination performance with an AUC of 0.88 and significance $p < 0.001$.

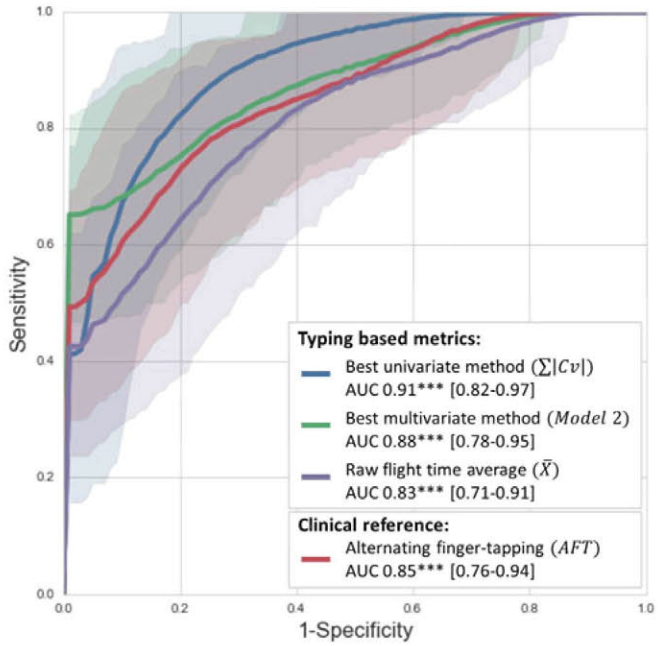

Fig. 4. Comparison of receiver operating characteristic (ROC) curves showing the classification rate for the typing based metrics, including raw flight time average ($\bar{X}$) and best performing univariate ($\sum |Cv|$) and multivariate methods ($Model2$), with the alternating finger-tapping test AFT. Statistical significance of the MannWhitney U test is estimated to reject the null hypothesis that the two groups, PD and CNT, come from the same population. Statistical significance noted as: $p < 0.001$(***), $p < 0.01$(**) and $p < 0.05$(*).

is run using the best model settings estimated in the inner loop and storing the score for the left-out sample.

We used two tests to evaluate the ability of each metric, typing features and models' scores, to correctly separate the referred classes: the Receiver Operating Characteristic (ROC) analysis and the Mann-Whitney U test to reject the null hypothesis that the controls and the Parkinson's samples come from the same distribution.

The ROC analysis consists of an iterative process that monotonically increases the value of the metric under study to define a dynamic threshold. On each iteration the current threshold value is evaluated as a binary classifier that separates Parkinson's and controls. The output is a set of sensitivity/(1-specificity) pairs that are joined to draw the corresponding ROC curve. The Area Under the Curve (AUC) can be interpreted as the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. Moreover, this metric allows a reliable comparison of the performance even when the number of cases and controls is not fully balanced, as it is the case of our study dataset (48% PD, 52% CNT). A sampling with replacement method (1,000 bootstraps) defines a ROC distribution from which we compute the average and confidence intervals of the AUC values to describe the classification performance of each metric.

III. RESULTS

Table II shows the results obtained for the univariate feature evaluation. The values presented for each typing metric include: mean value and standard deviation grouped by condition, average Area Under the Curve (AUC) for the bootstrapped ROC distribution, AUC confidence interval computed as the $[5^{th}, 95^{th}]$ percentiles on the resulting AUC values, and the Mann-Whitney significance test outcome.

Table III summarizes the results of the multivariate analysis. We evaluate nine different models defined as the possible combinations of three feature selectors and three classifiers. These methods were selected to represent different families of common machine learning approaches. Feature ranking methods used are Lasso, Ridge regression and Gini impurity based random forests used as estimators in a recursive feature elimination framework. Classification methods considered are logistic regression, linear kernel Support Vector Machines (SVM) and AdaBoost.

Alternating finger-tapping (AFT), a quantitative upper limb motor test commonly used in clinical trials to monitor PD signs,

TABLE IV
METHODS COMPARISON

Method	Avg. (std) Parkinson's	Avg. (std) Control	AUC [5%,95%]	Specificity	Sensitivity	Significance Adjusted	Significance Unadjusted
Best Univariate ($\sum \lvert Cv \rvert$)	3.290 (1.071)	1.839 (0.656)	0.91 [0.82, 0.97]	0.81	0.81	$p < 0.001$	$p = 0.002$
Best Multivariate ($Model2$)	0.825 (1.109)	−0.663 (0.730)	0.88 [0.78, 0.95]	0.84	0.73	$p < 0.001$	$p = 0.002$
Raw Flight Time Average ($\bar{X}$) (s)	0.870 (0.283)	0.566 (0.155)	0.83 [0.71, 0.91]	0.72	0.73	$p < 0.001$	$p = 0.003$
Alternating finger-tapping (AFT)	49.17 (10.65)	67.54 (14.11)	0.85 [0.76, 0.94]	0.78	0.75	$p < 0.001$	$p = 0.002$

The table compares the performance of the touchscreen typing based metrics, including the raw flight time average ($\bar{X}$) and best performing univariate ($\sum \lvert Cv \rvert$) and multivariate methods ($Model2$), with the alternating finger-tapping test AFT. The presented methods improve the discrimination ability of the reference test (AFT: 0.85[0.76, 0.94] AUC and 0.75/0.78 sensitivity/specificity), with 0.91[0.82, 0.97] AUC and 0.81/0.81 sensitivity/specificity for the best performing feature ($\sum \lvert Cv \rvert$) and 0.88[0.78, 0.95] AUC and 0.73/0.84 sensitivity/specificity for the best multivariate model ($Model2$). The adequacy of the proposed methods to enhance the differences of the typing patterns between Parkinson's subjects and controls is stressed by the comparison with the raw signal based metric ($\bar{X}$: 0.83[0.71, 0.91] AUC and 0.73/0.72 sensitivity/specificity). The presented sensitivity/specificity pairs correspond to the closest-to-(0,1) cut-off point. The unadjusted statistical significance is computed with two-sided Mann-Whitney U test. The adjusted significance tests were computed with logistic regression models including gender and age as co-variates. For the developed methods none of the co-variates reached statistical significance.

is used as the reference metric to evaluate the performance of the proposed method. Also, we include the average of the unprocessed flight time signal ($\bar{X}$) as a starting point to show the improvement introduced by our solution to the discrimination ability measured on the raw typing data. We replicate the evaluation framework used in our methods to test the classification performance of these two reference metrics in our cohort.

Fig. 4 and Table IV show the performance comparison of the touchscreen typing based metrics, i.e. raw flight time average and the developed methods, with the AFT test reference. Raw flight time average ($\bar{X}$) presents an AUC of 0.83 [0.71-0.91]. The best performing typing feature, covariance sum ($\sum |Cv|$), presents an AUC of 0.91 [0.82-0.97]. The best multivariate method ($Model2$), a combination of L1-regularized feature selection plus a linear SVM as the final estimator, scores an AUC of 0.88 [0.78-0.95]. AFT test performance measured in our cohort achieves an AUC of 0.85 [0.76-0.94]. Sensitivity and specificity values shown in Table IV are estimated using the closest-to-(0,1) criterion to define the cut-off point [23]. Unadjusted p-values present the results of the Mann-Whitney U test to reject the null hypothesis that PD and CNT subjects come from the same population. Adjusted significance tests the null hypothesis that the metric under scrutiny does not contribute to the separation between PD and control groups in a logistic regression model accounting for sex and age.

Finally, we evaluated the classification performance of the proposed methods for different signal lengths, in order to analyze the appropriate duration of continuous typing that would be necessary to achieve significant results. In Fig. 5 we illustrate the results of this analysis for our best univariate and multivariate methods.

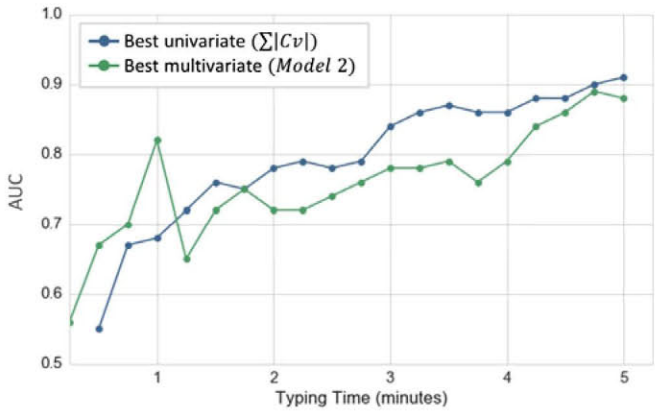

Fig. 5. The figure shows the evolution of the area under the ROC curve (AUC) for the best performing feature ($\sum |Cv|$) and best multivariate model ($Model2$) as we increment the amount of typing data used to perform the analysis. We observe a clear improvement of the classification performance as the duration of the analyzed typing series increases.

IV. DISCUSSION

In this work we propose an algorithm to identify PD motor signs by analyzing the typing activity on smartphones independently of the typed text. Users do not need to wear any sensor or remember to perform a structured test. Compliance depends only on the act of installing the software. Once installed, data collection happens automatically without interfering with the normal use of the device.

The current clinical standard used to quantify PD stage and progress present some limitations that define a clear need in the treatment and control of the disease. This scenario has led to the study and development of different alternatives in attempting to complete and complement UPDRS information.

In our cohort, a commonly used quantitative method that evaluates upper limbs dexterity, alternating finger-tapping (AFT), discriminated both populations with an AUC of 0.85 with 0.75/0.78 sensitivity/specificity. The proposed methods outperform this clinical reference, achieving an AUC of 0.91 with 0.81/0.81 for the best performing typing feature, the covariance sum ($\sum |Cv|$), and an AUC of 0.88 with 0.73/0.84 for the best multivariate method, a pipeline comprised of L1-regularized feature selection and a linear SVM as the final classifier.

We believe that our approach is able to achieve such performance because of bradykinesia, bimanual coordination problems and other PD signs that may alter typing kinetics in a way detectable through a keystroke timing data analysis. PD motor impairment, in the particular case of the FT signal, may impede

PD patients to press and release the keys in a consistent manner, which we hypothesize would induce irregular flight times (similar to what may be seen in finger tapping tests). Our results are consistent with that hypothesis in that the typing signal distribution for PD patients has a greater dispersion and temporal variability, i.e. the heteroscedasticity measured using those features. The improvement achieved in the classification rate, compared to the alternating finger-tapping test, may be due to the fact that our features (i.e. skewness, kurtosis and covariance of the FT distribution) have been carefully defined to specifically capture these motor abnormalities that are a direct representation of PD signs. The approach of constructing the typing signal as a sequence of consecutive signal segments allows an intra-subject analysis, which optimizes the detection of the internal variability introduced by PD signs.

One of the main difficulties when using the typing signal as the unique source of information is the risk of measuring variables that are representative of the typing style but do not capture the effect of PD signs. We limit this effect by applying a normalization phase that forces a zero mean. This focuses the analysis on the FT distribution shape and variability. Another external factor that has to be taken into account when studying the potential limitations of the proposed method is the requirement of a minimal number of signal samples to make the sub-distribution analysis consistent. To collect enough information from the natural typing signal, a minimum level of skills in touchscreen typing is demanded in order to meet the established criteria. We consider that, taking into account the rapid growing rate of smartphone users, typing skills will not limit the application of this method.

Our methods were validated in a controlled environment. Participants were asked to type for 5 minutes to complete the touchscreen test. Although they were instructed to type as they would normally do in order to reflect actual routine use of the device, further analysis will be necessary to discard a significant influence of the controlled test on their typing behavior. Regarding the 5-minute duration of the test, we understand that not all the smartphone users are likely to continuously type for this amount of time, however, the proposed methodology can be applied on natural typing signals collected for longer periods of time whose aggregate active typing time is 5 minutes or more.

This study is a step towards the final goal of developing an automated biometric tool for diagnostic and therapeutic decision support in PD. The presented methodology compares well to standard clinically-used methods in terms of its ability to differentiate PD participants from controls, and is able to do so from information collected from touchscreen typing activity. In our cohort, PD population presented mild signs (average UPDRS-III score of 17.76 ± 7.92 and range [6, 41]), this suggests that the proposed features are able to discriminate PD from controls even at early disease stages. However, as a pilot study, the findings of this research must be considered with caution. A further validation of this methodology would require a larger and better balanced cohort that enables a comprehensive review of the influence of the potential confounding variables mentioned in this work and others, such as medication state and cognitive deficits. Future work will include new studies to collect subjects' daily interaction with their smartphones in order to validate the applicability of the presented methods in a passively-monitored environment. Additional information from user's daily interaction with smartphones, such as pressure, gesture typing and accelerometer data, could be used to complement our keystroke based analysis. Method functionality could also be improved with the appropriate algorithms. Turning from a classification to a regression model, it may be possible to quantify a continuous metric for the natural progression of the disease over continuous motor function evaluations.

V. Conclusion

An approach to a more continuous, objective, and convenient tool to quantify PD related motor impairment is presented in this paper. The method suggests that motor anomalies in PD can be detected through analysis of keystroke dynamics during typing on smartphone touchscreens. The computed typing metrics show significant changes across the different studied groups: PD participants and healthy controls. In terms of classification, the best performing typing feature presents a 0.91 AUC rate, and sensitivity/specificity 0.81/0.81. The best multivariate model scores 0.88 AUC and 0.73/0.84 sensitivity/specificity. The proposed methods are comparable or improve the performance of the a reference motor test (AFT) measured in our cohort, 0.85 AUC and 0.75/0.78 sensitivity/specificity. Based on the analysis of the routine typing signal, the proposed approach introduces a transparent way to evaluate the motor function. In future work, a clinical study will validate our technique in a larger cohort of patients and controls and for an extended period of time that captures additional potential confounding variables.

Acknowledgements

The authors would like to thank the Federación Española Parkinson and the M+Vision and MIT linQ faculty for their guidance in developing this project. And also would like to thank their many clinical collaborators at MGH in Boston, HM-CINAC, Hospital "12 de Octubre," and Hospital Clínico San Carlos in Madrid for their insightful contributions, and the developers of the open source AnySoftKeyboard project.

References

[1] L. M. L. de Lau and M. M. B. Breteler, "Epidemiology of Parkinson's disease," *Lancet. Neurol.*, vol. 5, no. 6, pp. 525–35, Jun. 2006.

[2] B. Thomas and M. F. Beal, "Parkinson's disease," *Human Mol. Genetics*, vol. 16, no. R2, pp. R183–R194, Oct. 2007.

[3] F. L. Campos *et al.*, "Rodent models of Parkinson's disease: Beyond the motor symptomatology," *Frontiers Behav. Neurosci.*, vol. 7, Jan. 2013, Art. no. 175.

[4] D. J. Brooks, "Optimizing Levodopa therapy for Parkinson's disease with levodopa/carbidopa/entacapone: Implications from a clinical and patient perspective," *Neuropsych. Dis. Treatment*, vol. 4, no. 1, pp. 39–47, Feb. 2008.

[5] 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: Official J. Movement Disorder Soc.*, vol. 18, no. 7, pp. 738–50, Jul. 2003.

[6] A. L. Taylor Tavares *et al.*, "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," *Movement Disorders: Official J. Movement Disorder Soc.*, vol. 20, no. 10, pp. 1286–98, Oct. 2005.

[7] A. J. Noyce *et al.*, "Bradykinesia-Akinesia incoordination test: Validating an online keyboard test of upper limb function," *PLoS One*, vol. 9, no. 4, Jan. 2014, Paper e96260.

[8] C. N. Homann *et al.*, "The Bradykinesia Akinesia Incoordination test (BRAIN TEST), an objective and user-friendly means to evaluate patients with parkinsonism," *Movement Disorders: Official J. Movement Disorder Soc.*, vol. 15, no. 4, pp. 641–647, Jul. 2000.

[9] P. Arias *et al.*, "Validity of the finger tapping test in Parkinson's disease, elderly and young healthy subjects: Is there a role for central fatigue?" *Clin. Neurophysiol.: Official J. Int. Fed. Clin. Neurophysiol.*, vol. 123, no. 10, pp. 2034–2041, Oct. 2012.

[10] A. Sánchez-Ferro *et al.*, "New methods for the assessment of Parkinson's disease (2005 to 2015): A systematic review," *Movement Disorders*, vol. 31, no. 9, pp. 1283–1292, Sep. 2016.

[11] J. A. Serrano *et al.*, "Participatory design in Parkinson's research with focus on the symptomatic domains to be measured," *J. Parkinson's Dis.*, vol. 5, pp. 187–196, 2015.

[12] J. Stamford, P. Schmidt, and K. Friedl, "What engineering technology could do for quality of life in Parkinson's disease: A review of current needs and opportunities," *IEEE J. Biomed. Health Inf.*, vol. 19, no. 6, pp. 1862–1872, Aug. 2015.

[13] W. Maetzler *et al.*, "Emerging therapies for gait disability and balance impairment: Promises and pitfalls," *Mo. Disorders: Official J. Movement Disorder Soc.*, vol. 28, no. 11, pp. 1576–1586, Sep. 2013.

[14] N. L. Keijsers *et al.*, "Detection and assessment of the severity of Levodopa-induced dyskinesia in patients with Parkinson's disease by neural networks," *Movement Disorders: Official J. Movement Disorder Soc.*, vol. 15, no. 6, pp. 1104–1111, Nov. 2000.

[15] N. L. W. Keijsers *et al.*, "Ambulatory motor assessment in Parkinson's disease," *Movement Disorders: Official J. Movement Disorder Soc.*, vol. 21, no. 1, pp. 34–44, Jan. 2006.

[16] J. Cancela *et al.*, "Feasibility study of a wearable system based on a wireless body area network for gait assessment in Parkinson's disease patients," *Sensors*, vol. 14, no. 3, pp. 4618–4633, Jan. 2014.

[17] A. Salarian *et al.*, "Ambulatory monitoring of physical activities in patients with Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 12, pp. 2296–2299, Dec. 2007.

[18] A. Tsanas *et al.*,"Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, May 2012.

[19] B. M. Bot *et al.*,"The mPower study, Parkinson's disease mobile data collected using ResearchKit," *Nature Sci. Data*, vol. 3, pp. 1–9, 2016.

[20] L. Giancardo *et al.*, "Psychomotor impairment detection via finger interactions with a computer keyboard during natural typing," *Sci. Rep.*, vol. 5, 2015, Art. no. 9678.

[21] L. Giancardo *et al.*, "Computer keyboard interaction as an indicator of early Parkinson's disease," *Sci. Rep.*, vol. 6, no. 6, Oct. 2016, Art. no. 34468.

[22] S. J. Sheather and M. C. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," vol. 53, pp. 683–690, 1991.

[23] N. J. Perkins and E. F. Schisterman, "The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve," *Amer. J. Epidemiol.*, vol. 163, no. 7, pp. 670–675, Apr. 2006.

Authors', photographs and biographies not available at the time of publication.