UC San Diego
UC San Diego Previously Published Works

Title

Effects of mood and aging on keystroke dynamics metadata and their diurnal patterns in a large open-science sample: A BiAffect iOS study.

Permalink

https://escholarship.org/uc/item/0nc3d91k

Journal

Journal of the American Medical Informatics Association, 27(7)

Authors

Vesel, Claudia
Rashidisabet, Homa
Zulueta, John
et al.

Publication Date

2020-07-01

DOI

10.1093/jamia/ocaa057

Peer reviewed

AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

OXFORD

Research and Applications

Effects of mood and aging on keystroke dynamics metadata and their diurnal patterns in a large open-science sample: A BiAffect iOS study

Claudia Vesel,[1] Homa Rashidisabet,[1] John Zulueta,[2] Jonathan P. Stange,[2] Jennifer Duffecy,[2] Faraz Hussain,[2] Andrea Piscitello,[3] John Bark,[2] Scott A. Langenecker,[4] Shannon Young,[5] Erin Mounts,[5] Larsson Omberg,[5] Peter C. Nelson,[3] Raeanne C. Moore,[6] Dave Koziol, [7] Keith Bourne,[7] Casey C. Bennett,[8,9] Olusola Ajilore, [2] Alexander P. Demos,[10] and Alex Leow[1,2,3]

[1]Department of Bioengineering, University of Illinois at Chicago, Chicago, Illinois, USA, [2]Department of Psychiatry, University of Illinois at Chicago, Chicago, Illinois, USA, [3]Department of Computer Science, University of Illinois at Chicago, Chicago, Illinois, USA, [4]Department of Psychiatry, University of Utah, Salt Lake City, Utah, USA, [5]Sage Bionetworks, Seattle, Washington, USA, [6]Department of Psychiatry, University of California, San Diego, San Diego, California, USA, [7]Arbormoon Software, Inc, Ann Arbor, Michigan, USA, [8]College of Computing and Digital Media, DePaul University, Chicago, Illinois, USA, [9]School of Intelligence, Hanyang University, Seoul, Korea and [10]Department of Psychology, University of Illinois at Chicago, Chicago, Illinois, USA

Corresponding Author: Alex Leow, MD, PhD, Department of Psychiatry, University of Illinois at Chicago, 1601 W. Taylor St., SPHPI MC 912, Chicago, IL 60612, USA; alexfeuillet@gmail.com

ABSTRACT

Objective: Ubiquitous technologies can be leveraged to construct ecologically relevant metrics that complement traditional psychological assessments. This study aims to determine the feasibility of smartphone-derived real-world keyboard metadata to serve as digital biomarkers of mood.

Materials and Methods: BiAffect, a real-world observation study based on a freely available iPhone app, allowed the unobtrusive collection of typing metadata through a custom virtual keyboard that replaces the default keyboard. User demographics and self-reports for depression severity (Patient Health Questionnaire-8) were also collected. Using >14 million keypresses from 250 users who reported demographic information and a subset of 147 users who additionally completed at least 1 Patient Health Questionnaire, we employed hierarchical growth curve mixed-effects models to capture the effects of mood, demographics, and time of day on keyboard metadata.

Results: We analyzed 86 541 typing sessions associated with a total of 543 Patient Health Questionnaires. Results showed that more severe depression relates to more variable typing speed ($P < .001$), shorter session duration ($P < .001$), and lower accuracy ($P < .05$). Additionally, typing speed and variability exhibit a diurnal pattern, being fastest and least variable at midday. Older users exhibit slower and more variable typing, as well as more pronounced slowing in the evening. The effects of aging and time of day did not impact the relationship of mood to typing variables and were recapitulated in the 250-user group.

Conclusions: Keystroke dynamics, unobtrusively collected in the real world, are significantly associated with mood despite diurnal patterns and effects of age, and thus could serve as a foundation for constructing digital biomarkers.

Key words: keystroke dynamics, mHealth, mood, smartphone applications in health

INTRODUCTION

Background and significance

Traditional ways of assessing mental health usually rely on clinical evaluations or diagnostic interviews or mood rating scales that are intermittently administered in a controlled environment (ie, a clinic or laboratory). In addition, such measures frequently depend on patients' self-reports or reports from their family members or caretakers, and thus are subject to recall and recency biases, especially in individuals with impaired insight or cognition.[1] They may also take extended amounts of time and be expensive. Moreover, the intermittent "snapshot" nature of these traditional assessment models may not reflect intraindividual variability (IIV), which could serve as relevant additional clinical information. For example, the time of the day an assessment is administered[2] and the sleep quality of the patient[3] have been shown to affect cognitive performance. Therefore, clinical assessment can be enhanced and extended by granular, time-dependent metrics captured in ecologically relevant daily tasks in real-world settings.[4,5]

Current advances in smartphones and wearable devices can be leveraged for real-world continuous and unobtrusive data collection, possibly granular enough to construct features that could capture cognitive dynamics and potentially detect subtle intra- and interindividual heterogeneity.[6–10] Thanks to their high temporal resolution and the volume of sessions available, these approaches lend themselves particularly well to the study of diurnal patterns. In this study, we show how real-world data collected from smartphone keyboard in a large-scale open-science project can provide clinically relevant information. Specifically, we assess IIV through smartphone typing behaviors collected when users normally and routinely engage with their smartphones throughout the day as it relates to self-reported mood.

IIV can manifest as the inconsistency or fluctuations in task performance within an individual throughout the day.[11] Disruptions in circadian rhythms may drive early neurodegeneration,[12] suggesting that the relationship between diurnal patterns and IIV may be a sensitive marker of early cognitive decline, as supported by 2 recent studies[13,14] investigating typing speed collected on a virtual keyboard of a tablet and a physical keyboard.

Circadian rhythms are also clinically relevant for the onset, symptomatology, and trajectory of mood disorders.[15–17] Studies using data collected from wearable devices have proposed diurnal phase shifts as potential markers for mood disorders.[18] Further, inherent to mood disorders is the chronic course of illness, at times relapsing and remitting over the course of weeks, days, or even hours,[19] thus supporting the advantage of more frequent, dense sampling as evidenced in several studies.[20–22] For example, one recent study leveraged various sensors from wearable devices to predict mood states for the next 3 days and infer depressive, manic, or hypomanic episodes in individuals suffering from major depressive disorder and bipolar disorder.[23]

Objective

This study aims to investigate the effects of mood, age, and diurnal patterns on naturalistic real-world typing acquired using a mobile health (mHealth) smartphone technology in a citizen science sample. We hypothesize that (1) keyboard typing features will be significantly associated with mood ratings over time and (2) the association with mood may be modulated by effects of age and diurnal changes within days.

MATERIALS AND METHODS

Study design and demographics

Our research was structured around an mHealth application, BiAffect (https://www.biaffect.com/), that is freely available on Apple's app store and open to all U.S. adults for study enrollment directly via their iPhones. Once enrolled, BiAffect replaces the standard iOS keyboard with a cosmetically similar keyboard and records keystroke dynamics metadata regardless of whether the user is texting, writing an email, posting on social media, etc., while preserving anonymity of content (ie, not what one types but how it is typed). The application is supported by Apple's open source ResearchKit framework (Figure 1).[24]

Onboarding and data collection

Data analyzed in this study were collected within the first 15 months of our ongoing BiAffect study, comprising the most active users of our custom keyboard who also chose to provide their demographic information. In addition to keystroke dynamics, participants were prompted and encouraged to report their depression symptom severity using the Patient Health Questionnaire-8 (PHQ-8) (omitting the suicidality question)[25] once a week. The PHQ is a valid and reliable measure that has been used extensively in clinical care research[26–28] and shown to be sensitive to indication of depressive episodes and measurement of treatment response.[29–33]

The study was approved by the University of Illinois at Chicago's Institutional Review Board and all participants provided an informed e-consent on the BiAffect app, designed in collaboration with Sage Bionetworks (Seattle, WA) (see Supplementary Figures 1 and 2 for more details).

Data processing

The BiAffect keyboard records timestamps of keypress events and their type (ie, character, punctuation, backspace, autocorrect, or autosuggestion). Operationally, one typing session is initiated when the keyboard is activated, and terminated after 8 seconds of inactivity, or at the time of keyboard deactivation. Session duration is defined as the length of a typing session in seconds.

Examining the distribution of interkey delays to understand typing speed, pausing, and typing variability

To analyze the speed of typing during a session, we followed similar approaches to those published in studies from our group[34,35] and others.[13] First, interkey delay (IKD) was calculated as the time difference between 2 consecutive keypresses and we restricted our analysis to IKDs between character-to-character keypress events only. Other transitions were excluded because they may encode additional cognitive processes beyond typing alone (eg, character-to-backspace may encode processes subserving self-monitoring of typing mistakes and their correction). Further, as IKDs exhibit a heavy right-tailed distribution (Figure 1) with values between 0 and 8 seconds (longest delay allowed in a session), we hypothesized that long IKDs encode events other than word-level typing behavior (eg, pauses), while short IKIDs would be more relevant for capturing the everyday casual typing speed. Thus, the typing speed of a typing session is operationally inferred using the median (50th percentile) IKD in that session (the median was selected as it is the most stable estimate of central tendency for positively skewed distributions). By contrast, pausing is operationally defined as the 95th percentile of the session-level IKD distribution (the 95th percentile is chosen as it
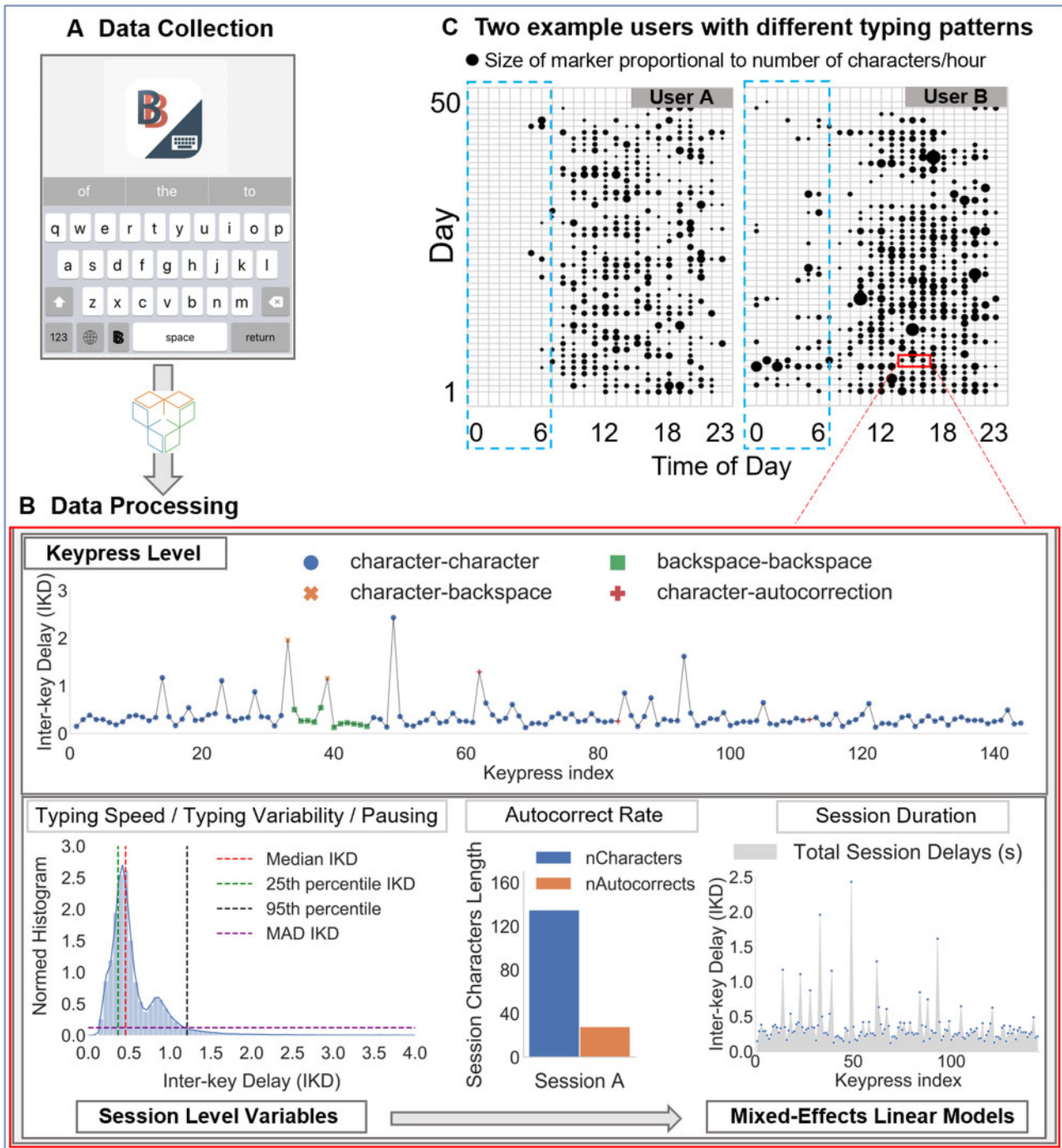
Figure 1. Overview of BiAffect data collection and feature extraction process. (A) Keypress-level typing metadata are collected via the BiAffect keyboard and stored by Sage Bionetworks. (B) Interkey delays (IKDs) for keypress transitions from character to character are aggregated at a session level to obtain the 25th, 50th (median), and 95th percentile IKDs and the median absolute deviance (MAD) IKD. Typing accuracy in a session is defined as the autocorrect rate (ie, ratio of autocorrect instances to the total number of characters). Last, the session duration (seconds) in each session is obtained by aggregating all delays in a session. (C) An example for the hourly typing activity over multiple days from 2 active users is shown in the top right as an illustration of the potential patterns captured via continuous, unobtrusive collection. The blue dashed line highlights the different levels of activity at night, with user B exhibiting a more irregular activity pattern than user A. Size of the marker is proportional to the number of characters typed per hour.

is commonly employed to capture behavior that is outside of the norm; also see Supplementary Table 14). Last, in order to quantify the typing variability at a session level, we used the median absolute deviance (MAD) of IKDs.

Typing accuracy

To infer typing accuracy, the ratio of autocorrect instances relative to the total session character count was recorded (number autocorrects/total number of characters per session).

Typing mode (1- vs. 2-handed typing)
Careful consideration was given to the typing mode (using 1 or 2 hands when typing), as it is likely to affect typing speed. To classify typing mode, we developed and validated an approach using linear regression. This method was validated via independent test data for over 220 sessions (40 182 total keypresses) collected on internal testing phones, yielding >99% accuracy (see Supplementary Figures 3 and 4).

Time of day
In order to assess diurnal patterns, every session was stamped with the hour of the day in the user's local time and then aggregated in 3-hour increments, yielding 8 time points throughout the day. Data were aggregated in this fashion to allow more stable estimates of the parameters extracted from the metadata.

Eight-Item PHQ
The standard instrument has 9 items and asks participants to report how bothered they have been by items over the past 2 weeks on a scale from 0 (*not at all*) to 3 (nearly every day). Given our open science design, which does not permit clinical monitoring or oversight, we chose to eliminate the suicide item from our study. The resulting PHQ-8[25] had total scores ranging from 0 to 24, with higher scores indicating greater depressive symptoms. Additionally, as the PHQ-8 is designed to measure depression severity over the course of 2 weeks, we decided to prompt users, on a weekly basis instead of every 2 weeks to counter common attrition in self-reports, to perform mood assessments using PHQ-8.

The relationship between depression symptom severity and keystroke dynamics was analyzed by propagating the PHQ-8 score to typing sessions that occurred within a time window of a given PHQ-8 score. First, the very first PHQ-8 rating in every user was used to tag all typing sessions within 2 weeks prior to this rating (as the PHQ is designed to capture overall depression severity within the preceding 2 weeks of a rating). When multiple PHQs were available, a linear interpolation was then performed for the typing sessions in between consecutive self-reports if (1) their score difference was lower than 5 and (2) PHQ-8 scores were not further than 2 weeks apart. Otherwise, for each pair of consecutive PHQs that were temporally disconnected or exhibiting a large score difference, we propagated the second PHQ-8 score 2 weeks prior to it (in the case of temporal disconnection), or up to the day after the preceding PHQ-8 (in the case of a large score difference). Last, for any typing sessions between each pair of consecutive PHQs that remained untagged, we explored 3 different schemes that carry the score of the first PHQ-8 of this pair after its rating by up to 0, 1, and 2 weeks. Similarly, these 3 variations were employed to carry the score of the very last PHQ-8 of each user after its rating (see Supplementary Tables 1-5).

Statistical analysis
We conducted growth curve mixed-effects (multilevel) models[36] in R (version 3.6.1; R Foundation for Statistical Computing, Vienna, Austria) and lme4 (version 1.1-21)[37] using maximum likelihood fitting to examine dependent variables (DVs) of session-level typing speed, typing variability, typing accuracy, and session duration and their relationship to other session-level features and demographics. In total, the number of typing sessions that entered our 2-level mixed-effects models are n = 142 202 nested in the 250-user group (sample A) and N = 86 541 nested in the 147-user group (sample B;

a subset of sample A, sample B differs from A in that each of these 86 541 typing sessions can be linked back to a specific PHQ-8, thus allowing us to investigate the effect of mood on typing; see Figure 2 and Table 1).

Random effects
Random effects of the models included the user as the cluster in which sessions were nested, and we allowed each user to have their own intercept. This means that every subject could have their own mean of the point estimate of the DVs per session, thus allowing us to control for variance differences between users that cannot be later attributed to a fixed effect in our model. Further, this produces a more stable and generalizable point estimate per user. We note that our final models do not include the effects of gender or the phone type, as they explained no variance beyond the random intercept term for each user (Supplementary Tables 11 and 12).

The intraclass correlation coefficient (ICC) measures the homogeneity of the cluster (ie, the user) on the DV[38] and is calculated for our 2-level model intercepts only (null model) as the ratio of between cluster variance to total variance. These numbers can be interpreted as the cohesion or correlation within the cluster. We computed ICCs for models with DVs of 50th percentile IKD (ICC = 0.79), 95th percentile IKD (ICC = 0.35), MAD IKD (ICC = 0.55), session duration (ICC = 0.21), and the rate of autocorrect per session (ICC = 0.10). These ICC values all suggest that we do need to use 2-level model to examine our fixed effect. We further allowed each user to have their own orthogonal linear and quadratic slopes for the time of day and different intercepts for their typing mode. This allows more stable estimates of both typing mode and time of day, as these are essentially repeated measures within user (cluster) and allow for each user to potentially have their own unique difference in typing mode and effects over the course of the day. When we add fixed predictors to this multilevel model, it allows us to state whether there are reliable slopes across users. Because of the way they fit random and fixed slopes at the user level (cluster), these models are generally understood to be more generalizable potentially outside our individual samples.[36]

Fixed effects
The fixed-effects structure was tested in 2 ways (α was set to .05). First, fixed effects were constructed hierarchically through the addition of increasingly complex terms, and likelihood ratio tests (−2 log-likelihood of model fits) were used to compare nested models (against chi-square distribution) to ensure that increasingly complex terms were warranted (hierarchical growth models are detailed in Supplementary Tables 6-10). A −2 log-likelihood (ie, deviance) was used to assess model fit because of both the nested (ie, multilevel) nature of the data (users have multiple sessions) and because we allowed random slopes for users based on their typing mode and the time of the day that they were typing. Both the multilevels (sessions nested in users) and the random slopes make traditional metrics of R^2 and RSME from nonmultilevel regressions difficult to interpret and we report deviance because it can compare nested models and for consistency.[36,42] Second, the significance of individual predictors was assessed via estimating the degrees of freedom with Satterthwaite approximations.[39] The goal of this 2-step significance testing was to first ensure improvement in the overall model fit from a less complex model, and second to ensure the reliability of the individual parameters of this best fit model. Models involving IKD were tested hierarchically adding the
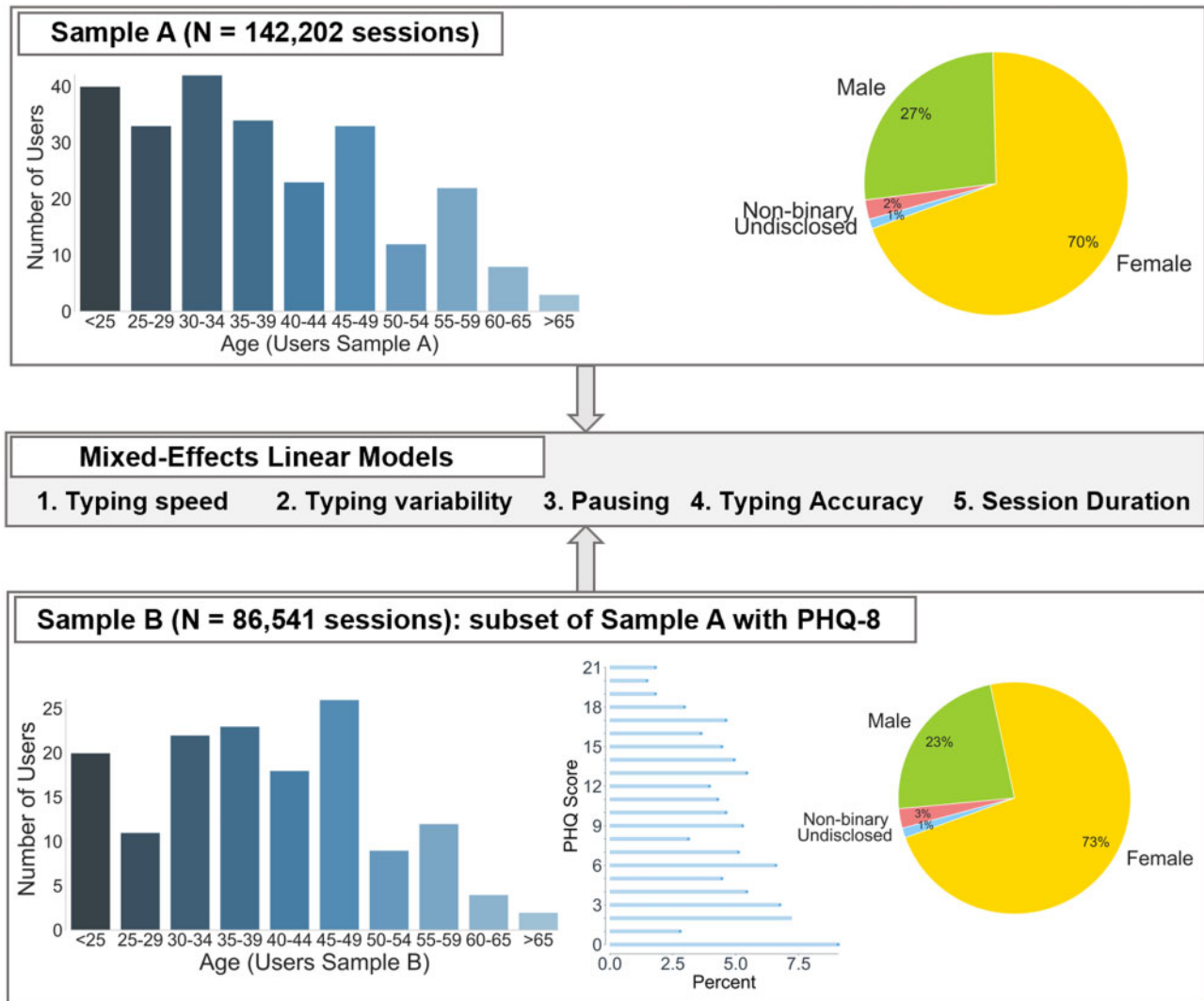
Figure 2. Overview demographics for the 2 samples used in mixed-effects linear models with dependent variables for typing speed, typing variability, pausing, typing accuracy, and session duration. Sample A consisted of 250 users (n = 142,202 sessions) and was used to investigate effects of time of day, age, typing mode (ie, 1- or 2-handed), total number of characters, etc. in a session. Sample B (147 users, n = 86,541 sessions) was a subset of sample A comprising users who completed at least 1 Patient Health Questonnaire-8 (PHQ-8) and was used to additionally investigate the effect of mood (543 total PHQs).

time of day, age and gender of the user, typing mode, and number of characters per session. Time of day and the total character length per session were entered as second-order orthogonal polynomials,[40] which allowed the independent assessment of significance of terms (linear and quadratic). Last, in order to quantify the effect of depressive symptoms, the PHQ-8 score was added as a linear fixed effect.

RESULTS

Of the 369 most active users (top 60%) chosen based on a minimum of 1000 keypresses per person, we analyzed typing sessions from 250 (sample A: median 12 151 [interquartile range (IQR), 4 482, 35 663], mean 30 544.9 ± 89 679 keypresses per person) who reported their demographic information (mean 37.75 ± 12.25 years of age; gender: 70% women, 27% men, 2% nonbinary, 1% undisclosed). Collectively, sample A comprises 142 202 typing sessions (median 181 [IQR, 74, 552], mean 568.8 ± 1508.9 sessions per user) that entered our mixed-effect models.

Additionally, in order to investigate the effects of mood on typing, we further identified a subset of 147 users (sample B: mean 39.53 ± 11.80 years of age, gender: 73% women, 23% men, 3% non-binary, 1% undisclosed), who completed at least 1 PHQ-8 (median 2 [IQR, 1, 5], mean 4.22 ± 5.90 PHQs per person; a total of 543 PHQs with a mean score of 9.0 ± 5.94). In total, 86 541 typing sessions entered the mixed-effects models for sample B (based on tagging typing sessions up to 2 weeks after the last PHQ-8; the main results were stable regardless of whether we propagated the last PHQ-8 rating 0, 1, and 2 weeks forward; see Supplementary Appendix for more details).

An overview of the keyboard data and demographics in the 250-user group and the 147-user group can be found in Table 1 and Figure 2.

Hierarchical growth-curve mixed-effects models were employed on the session level for typing speed (50th percentile IKD), pausing (95th percentile IKD), variability (MAD IKD), and accuracy (rate of autocorrect per session), as well as session duration (in seconds). All the trends observed in the larger model (250 users) hold true in the subsample (147 users) as well.

Table 1. Summary demographics, PHQ-8, and typing data distribution for sample A (142 202 sessions, 250 users) and sample B (86 541 sessions, 147 users)

	Sample A	Sample B
Age		
Mean ± SD	37.75 ± 12.25	39.53 ± 11.80
Gender		
% Women	70	73
% Male	27	23
% Nonbinary	2	3
% Undisclosed	1	1
Typing data		
Total number of typing sessions	142 202	86 541
Total number of users	250	147
Typing sessions per user		
Median (IQR)	181 (74-552)	207 (84-577)
Mean ± SD	568.8 ± 1508.9	588.71 ± 1133.4
Total keypresses per user		
Median (IQR)	12 151 (4482-35 663)	13 405 (4967-37 562)
Mean ± SD	30 544.9 ± 89 679	31 611.1 ± 66 054
Days of keyboard activity per user		
Median (IQR)	35 (6-15)	16 (9-29)
Mean ± SD	32.01 (54.51)	29.11 (45.34)
Self-reported psychiatric disorders[a]		
Bipolar disorder[b]	82/175	51/107
Depression	124/181	75/111
Anxiety	116/181	71/111
Attention deficit/hyperactivity disorder	46/181	29/111
Posttraumatic stress disorder	45/181	31/111
PHQ-8		
Total number self-reports	—	543
PHQ-8 score per person		
Median (IQR)	—	9.8 (5.3, 13.4)
Mean ± SD		9.0 (5.94)
PHQs per person		
Median (IQR)	—	2 (1-5)
Mean ± SD		4.22 ± 5.90

Sample B is a subset of A in that, in addition to demographic information, people in B also completed at least 1 PHQ-8.

IQR: interquartile range; PHQ-8: Patient Health Questonnaire-8.

[a]The denominator for each self-reported group is the total number of users in that sample (A or B) who completed the self-reported questionnaire for each psychiatric condition. The numerator represents the number of users who responded positive to a previous diagnosis for that condition.

bIncludes bipolar disorder I, bipolar disorder II, and other unspecified bipolar disorder.

Typing speed (50th percentile IKD), typing variability (MAD IKD), and pausing (95th percentile IKD)

Sample A

As shown in Table 2 (models 1-3A), we found a significant nonlinear diurnal pattern for typing speed (ie, 50th percentile IKD), typing variability (ie, MAD IKD), and pausing (ie, 95th percentile IKD), modeled using a second-order polynomial. People typed faster with less variability and exhibited the least amount of pausing midday (between noon and 3:00 PM in their respective time zones). We note a 11.36% (95% confidence interval [CI], 11.21%-11.52%) decrease in speed, a 12% (95% CI, 11.89%-12.13%) increase in variability, and a 12.2% (95% CI, 12.14%-12.23%) increase in pausing around midnight (between 0:00 and 2:00 AM) compared to midday.

There was a positive linear effect for age on median IKD, such that younger users (~20 years of age) typed 62% (95% CI, 61.2%-64.0%) faster, 57% (95% CI, 56.3%-58.9%) less variably, and paused 61% (95% CI, 59.8%-62.6%) less than older users (≥70 years of age). Further, we report a significant interaction between diurnal patterns and age, such that older people exhibited a more pro-nounced slowing in their typing speed and an increase in variability toward the end of the day (Figure 3).

Typing speed also exhibited a nonlinear relationship with the number of characters per session, with shorter sessions (20 characters) being 12.3% (95% CI, 11.85%-12.7%) slower and 12.4% (95% CI, 11.94%-12.9%) more variable than longer sessions (~120 characters). As expected, 1-handed typing sessions were found to exhibit slower and less variable typing speed than 2-handed sessions.

Sample B

The pattern observed in models 1-3A (Table 2) for time of day, age, total character length per session, and typing mode is recapitulated in model 1-3B for individuals who completed the PHQ-8. Further, we found that, on average, persons with a PHQ-8 score of 20 corresponded to a 2.2% (95% CI, 2.13%-2.33%) shortening of the 50th percentile IKD and a 7.8% (95% CI, 7.52%-8.13%) increase in pausing (higher 95th percentile IKD) when compared with persons with a PHQ-8 score of 0. Similarly, typing sessions corresponding to

Table 2. Summary of the final mixed effects models showing estimates for dependent variables of typing speed (model 1), typing variability (model 2), and pausing (model 3)

	Typing speed (median IKD)		Typing variability (MAD IKD)		Pausing (95th percentile IKD)	
	Model 1A	Model 1B	Model 2A	Model 2B	Model 3A	Model 3B
Fixed effects						
Intercept	0.296[c]	0.307[c]	0.123[c]	0.126[c]	0.902[c]	0.917[c]
	(0.005)	(0.006)	(0.002)	(0.003)	(0.015)	(0.018)
Time[d]	1.544[c]	1.216[c]	0.713[c]	0.495[c]	5.742[c]	3.758[c]
	(0.263)	(0.315)	(0.134)	(0.146)	(1.089)	(1.021)
Time[e]	3.357[c]	2.850[c]	1.446[c]	1.178[c]	10.639[c]	8.827[c]
	(0.239)	(0.260)	(0.117)	(0.116)	(0.897)	(0.814)
Age	0.069[c]	0.065[c]	0.026[c]	0.024[c]	0.204[c]	0.194[c]
	(0.005)	(0.006)	(0.002)	(0.003)	(0.014)	(0.017)
Typing mode	0.016[c]	0.016[c]	0.005[c]	0.004[b]	0.007	0.007
	(0.002)	(0.003)	(0.001)	(0.001)	(0.007)	(0.008)
Session characters length[d]	−3.717[c]	−3.068[c]	−1.543[c]	−1.284[c]	−9.109[c]	−7.128[c]
	(0.054)	(0.056)	(0.040)	(0.042)	(0.457)	(0.466)
Session characters length[e]	1.643[c]	1.319[c]	0.738[c]	0.600[c]	3.732[c]	2.499[c]
	(0.052)	(0.053)	(0.038)	(0.040)	(0.437)	(0.445)
PHQ	—	−0.002[c]	—	0.001[b]	—	0.020[c]
		(0.001)		(0.000)		(0.004)
Time[d] × Age	0.522[a]	0.644[a]	0.196	0.290[a]	1.069	1.368
	(0.256)	(0.304)	(0.130)	(0.140)	(1.054)	(0.979)
Time[e] × Age	0.406	0.297	0.333[b]	0.218	0.376	0.192
	(0.233)	(0.253)	(0.114)	(0.113)	(0.873)	(0.794)
Random effects	Variance	Variance	Variance	Variance	Variance	Variance
User	(Intercept)	0.005	0.005	0.001	0.001	0.029	0.026
User	Typing Mode	0.001	0.001	0.000	0.000	0.004	0.004
User	Time[d]	12.069	11.202	2.502	1.907	59.048	29.717
User	Time[e]	9.154	6.747	1.640	0.941	38.557	17.810
Residual	0.003	0.003	0.001	0.002	0.058	0.062
Model fit						
AIC	−438 810.2	−262 557.8	−525 470.1	−311 836.0	73.295	5554.95
BIC	−438 672.1	−262 417.3	−525 332	−311 695.5	211.405	5695.47
Log-likelihood	219 419.1	131 293.9	262 749.1	155 933.1	−22.648	−2762.47
Observations	142 202	86 541	142 202	86 541	142 202	86,541

Standard errors are listed in in parentheses. Sample A (142 202 sessions, 250 users) and sample B (86 541 sessions, 147 users) had fixed effects for orthogonal linear and quadratic effects of time, age, typing mode, and session character length. Sample B adds Patient Health Questionnaire-8. Dependent variables were typing speed, variability, and pausing.

AIC: Akaike information criterion; BIC: Bayesian information criterion; IKD: interkey delay; MAD: median absolute deviance.

[a]$P < .05$.
[b]$P < .01$.
[c]$P < .001$.
[d]1st order coefficient.
[e]2nd order coefficient.

a PHQ-8 score of 20 were found to exhibit a 3.3% (95% CI, 3.14%-3.39%) increase in typing variability, possibly driven by the faster typing speed yet longer pauses (Figure 4).

Model improvement was computed for sample B by adding the PHQ-8 score as a main effect to the final model for typing speed (χ^2 = 14.8, $P < .001$, deviance = −262 592), typing variability (χ^2 = 9.18, $P < .01$, deviance = −311 871), and pausing (χ^2 = 22.9, $P < .001$, deviance = 103 879).

For completeness, additional IKD percentile models were also explored for typing speed (the 25th and 75th percentile IKD), and for alternative percentile measurements of pausing (80th, 85th, and 90th percentile IKD; see the Supplementary Appendix). The longer pausing effect in more severe depression can be observed starting at the 85th percentile IKD model and remaining statistically significant at both the 90th and the 95th percentile IKD model.

## Typing accuracy (rate of autocorrect instances per session)

Sample A
Typing accuracy (Table 3) exhibited a nonlinear relationship with the number of characters per session, with longer sessions showing a 14.1% (95% CI, 13.4%-14.8%) increase in the number of autocorrect instances. As expected, more autocorrect instances related to faster typing (shorter median IKD).

Sample B
Lower typing accuracy found in sessions with a higher number of characters and faster typing described in model A was observed in model B as well. In addition, we found that (compared with a PHQ-8 score of 0) a PHQ-8 score of 20 led to a 7.2% (95% CI, 6.84%-7.67%) increase in typing mistakes.
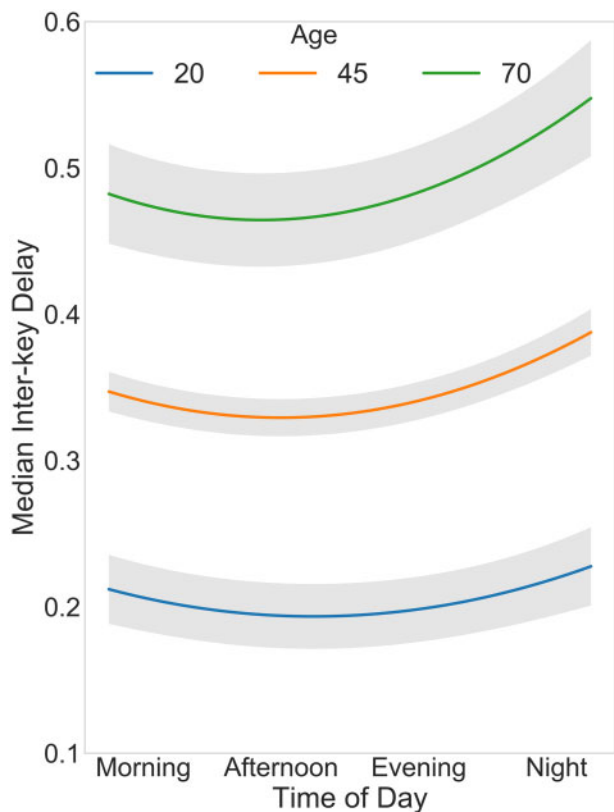
Figure 3. Linear mixed-effects model fit for median interkey delay (model 1, sample A in Table 2) as a function of age and time of day. A nonlinear diurnal pattern for typing speed was found, with the fastest typing occurring midday (12 PM to 3 PM) and the slowest typing at night. Older users are also shown to exhibit slower typing. Last, the significant interaction between age and time of day results in a more pronounced slowing in older users in the evening. Model obtained by plotting median interkey delay at different times in the day while fixing all other model covariates at their average value. The shaded area represents standard error ribbons.

Deviance testing was employed for the typing accuracy model after adding the PHQ-8 as a main effect to show model improvement ($\chi^2 = 6.12$, $P < .05$, deviance $= -378\,809$).

Session duration

Sample A

As shown in Table 3, we report that session duration was found to increase by 16.2% (95% CI, 13.4%-18.8%) in older people, even when accounting for the median IKD, which was previously found to increase with age (Table 2, models 1A and 1B). Further, we found a 13.0% (95% CI, 12.8%-13.3%) increase in session duration in sessions with a 0.2 autocorrect rate compared to those with 0 autocorrect events. In addition, we report a nonlinear diurnal pattern in session duration, with the shortest sessions occurring midday, coinciding with the time of day when people typed the fastest (Table 3, models 1A and 1 B and Figure 4). As expected, 1-handed sessions were found to be longer than 2-handed sessions.

Sample B

The same pattern was observed, with longer session duration observed in older users, in sessions with a lower typing accuracy, and in 1-handed typing sessions. In addition, we note that on average sessions corresponding to a PHQ-8 score of 20 were 7.95% (95%

CI, 7.73%-8.17%) shorter in duration when compared with sessions corresponding to a PHQ-8 score of 0, even when accounting for typing speed as a covariate, suggesting a decrease in overall phone use (Figure 4). The session duration model was found to improve by adding the PHQ-8 score as a main effect ($\chi^2 = 45.54$, $P < .001$, deviance $= 701\,787$).

Post hoc analysis for self-reported diagnoses as main effects after accounting for PHQ-8

Self-reported histories of depression, bipolar disorder, anxiety, attention-deficit/hyperactivity disorder, and posttraumatic stress disorder (Table 1) were tested as main effects in our models after accounting for PHQ-8 (sample B), and none were found to be significant after controlling for multiple comparisons (see the Supplementary Appendix for more details).

DISCUSSION

As part of a naturalistic crowd-sourced study on keystroke dynamics and mood in the real world, this study examined intra- and interindividual variability in keystroke dynamics metadata to reveal an underlying mood effect. To this end, we first identified various keystroke dynamics features, and established the feasibility of collecting large-volume, passively collected typing metadata using an open-science research model.

Keystroke dynamics relate to mood and age

In our previous published studies that analyzed the Android pilot data acquired in a much smaller sample of bipolar participants using a traditional research design,[34,35] we reported that keystroke dynamics features similar to those used in this paper, aggregated at a week level were able to predict depression severity at the end of that week (ie, the depression severity is the DV). By contrast, the main theme of the present article was to take a complementary approach by asking whether keystroke dynamics at the session level (thus at a much higher granularity) are modulated by mood ratings, as well as by other variables such as the time of day and user demographics.

Results established that typing speed exhibits slowing with age, while pausing between typing and variability in typing speed increase with age. Similarly, our hypothesis that keystroke dynamics features may relate to mood[34,35] is supported by our findings that in more severe depression there is a significantly higher variability in IKDs (as measured by MAD). This effect is likely due to longer pauses (the 95th percentile IKD within a session) yet slightly shorter 50th percentile IKD. This is consistent with reported findings of higher IIV in task performance in mood disorders.[41] Further, typing accuracy, as encoded using session-level autocorrect rates, also decreases in more depressed individuals. Last, sessions corresponding to elevated depressive symptoms were found to be shorter in duration, suggesting a decrease in smartphone keyboard use during more severe depression.

Collectively, these findings are in line with published literature that consistently demonstrated age-related cognitive decline and neurocognitive impairments in mood disorders for both average and deviation in cognitive performance,[41] and thus in support of our central theme of keystroke dynamics serving as digital biomarkers of mental health.
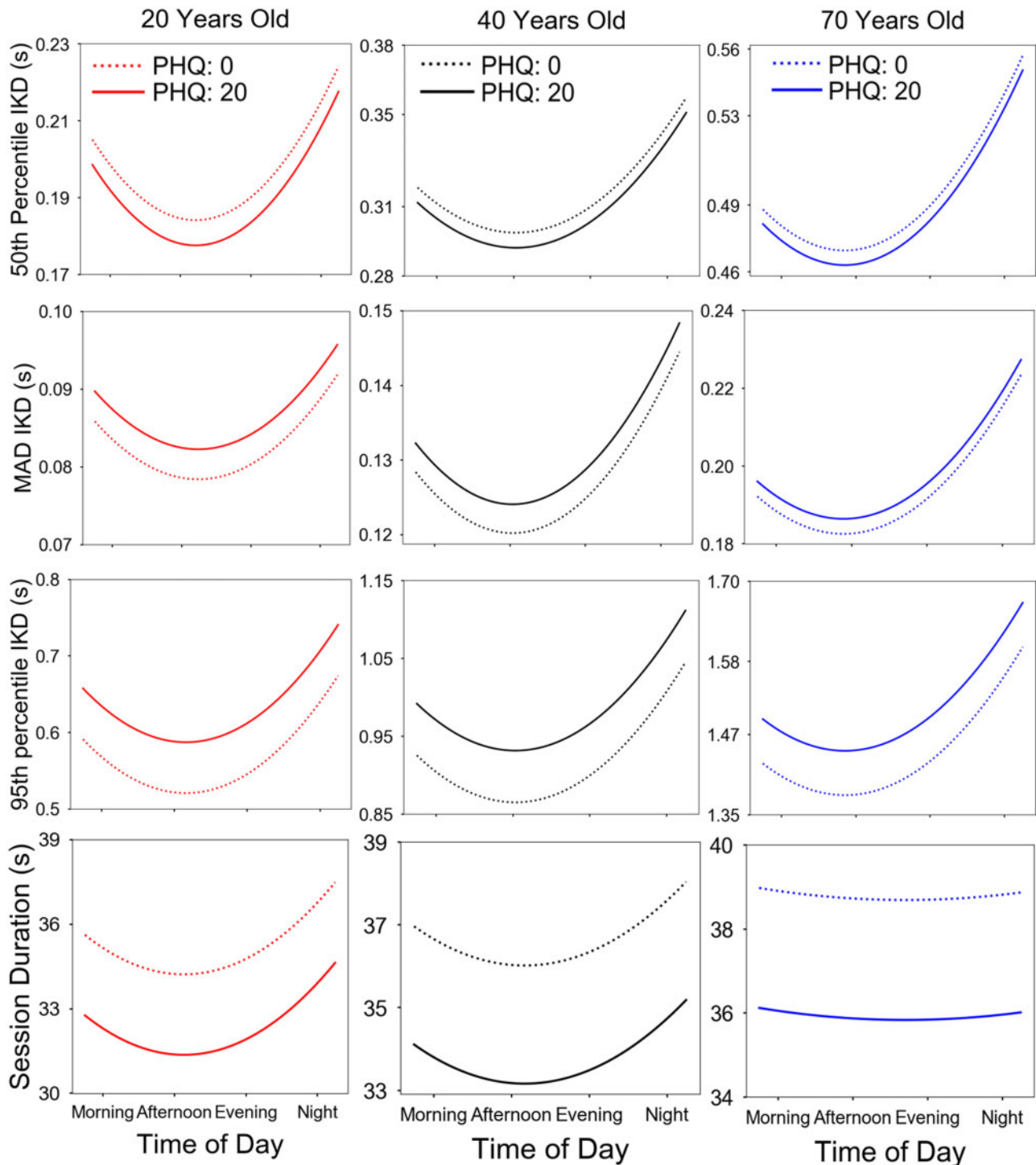
Figure 4. Linear mixed-effects model fits using sample B for 50th percentile interkey delay (IKD) (model 1), median absolute deviance (MAD) IKD (model 2), 95th percentile IKD (model 3), and session duration (model 5) as a function of age, time of day, and Patient Health Questonnaire-8 (PHQ) score. We found a nonlinear diurnal pattern for all dependent variables corresponding to fastest and least variable typing, shortest pauses, and shortest session duration midday (12-3 PM). Additionally, older users typed slower, typed more variably, paused longer, and had longer session durations than did young participants. Last, in more severe depression (higher PHQ score, continuous line), we found an increase in typing speed variability (higher MAD IKD), owing to faster typing (lower 50th percentile IKD), and longer pauses between typing (higher 95th percentile IKD), as well as shorter session durations. We note that the effects of PHQ remained consistent even when accounting for diurnal patterns and aging effects. Model obtained by plotting dependent variables at different times in the day while fixing all other model covariates at their average value.

Table 3. Summary of the final mixed-effects model showing estimates for dependent variables of typing accuracy (model 4) and session duration (model 5)

	Typing accuracy (rate autocorrect)		Session duration (seconds Z-scored)	
	Model 4A	Model 4B	Model 5A	Model 5B
Fixed effects				
Intercept	3.071^b	2.917^c	0.157^b	0.131^b
	(0.077)	(0.091)	(0.014)	(0.019)
Time^c	5.577	−1.243	0.650	0.920
	(7.618)	(7.624)	(1.202)	(1.200)
Time^d	0.808	−1.182	7.030^b	5.981^b
	(5.411)	(5.230)	(1.347)	(1.324)
Age	0.008	−0.050	0.045^b	0.047^a
	(0.075)	(0.088)	(0.013)	(0.018)
Typing mode	0.046	0.042	0.081^b	0.073^b
	(0.039)	(0.045)	(0.013)	(0.017)
Session characters length^c	39.752^b	33.722^b	221.491^b	186.201^b
	(2.942)	(2.901)	(0.795)	(0.770)
Session characters length^d	-36.109^b	-30.938^b	-29.392^b	-22.348^b
	(2.782)	(2.737)	(0.752)	(0.726)
Median IKD	-0.644^b	-0.611^b	0.276^b	0.274^b
	(0.014)	(0.018)	(0.004)	(0.005)
Rate Autocorrect	–	–	0.035^b	0.033^b
			(0.002)	(0.003)
PHQ	–	0.063^a	–	-0.044^b
		(0.025)		(0.007)
Time^c × Age	3.873	5.929	−1.557	−0.662
	(7.342)	(7.280)	(1.151)	(1.151)
Time^d × Age	0.447	−4.501	−1.824	−1.835
	(5.227)	(5.063)	(1.304)	(1.281)
Random effects	Variance	Variance	Variance	Variance
User | (Intercept)	1.366	1.125	0.038	0.045
User | Typing mode	0.147	0.108	0.023	0.022
User | Time^c	6792.687	4337.487	61.602	45.470
User | Time^d	1846.221	1072.357	104.535	68.311
Residual	7.523	7.284	0.549	0.512
Model fit				
AIC	691 935.0	418 293.16	319 344.961	188 420.291
BIC	692 083.06	418 443.05	319 502.801	188 579.553
Log-likelihood	−345 952.54	−209 130.58	−159 656.481	−94 193.145
Observations	142 202	86 541	142 202	86 541

Sample A (142 202 sessions, 250 users) and sample B (86 541 sessions, 147 users) had fixed effects for orthogonal linear and quadratic effects of time, age, typing mode, and session character length. Sample B adds the Patient Health Questionnaire-8 score. Model 5 also adds typing accuracy as a fixed effect. Dependent variables were typing accuracy and session duration.

AIC: Akaike information criterion; BIC: Bayesian information criterion; IKD: interkey delay; MAD: median absolute deviance.

[a]$P < .05$.
[b]$P < .001$.
[c] 1st order coefficient.
[d] 2nd order coefficient.

Keystroke dynamics further reflect circadian rhythm on a granular level

As our platform allows for the collection of high-volume repeated measures for each user, another major theme of our paper is establishing the feasibility of inferring highly granular intra-individual variations. Of particular importance for parsing through typing performance variability is the temporal richness of measurements that can capture diurnal patterns, suggestive of circadian rhythms affecting typing dynamics (slower and more variable typing at the end of a day). More importantly, our integrated approach unveiled complex effects, such as a time of day and age interaction, suggesting that older users exhibit a more pronounced slowing in their typing speed at night.

Limitations

Although our naturalistic approach offers ecological validity at scale when compared with studies employing traditional neuropsychological evaluations in artificial settings by allowing shorter and more frequent assessments, it also suffers from several limitations, such as the inability to verify self-reported diagnoses and symptom severity, as well as the lack of gold-standard clinician ratings. In addition, owing to the "in-the-wild" nature of the study, there are many unknown confounding factors that we are not able to control. Nevertheless, despite these shortcomings, our main conclusions are strongly significant and in line with other findings in published literature.

Further, mood self-reports (PHQ-8) were substantially sparser than keyboard data, so we elected to propagate their scores to typing

sessions within a time window to preserve as much typing data as possible. To this end, for pairs of consecutive PHQs not differing more than 5 points and no further apart than 2 weeks we used interpolation, although main results remain stable when alternative interpolation methods were used (Supplementary Material Tables 1-5).

Additionally, many enrolled participants (total of 989 users by June 2018) did not contribute enough keypresses to be included in our analyses, likely owing to BiAffect's custom keyboard not performing as well as the native iOS keyboard for them or possible privacy and confidentiality concerns. As the native iOS keyboard and the underlying technology are a trade secret, we thus cannot precisely quantify how the native keyboard differs from ours. However, one possible difference is in the autosuggest and autocorrect functionality, which in the native iOS keyboard is adaptive and thus is likely implemented using state-of-the-art natural language processing algorithms, whereas in the BiAffect keyboard, we employed a simple N-gram approach.

While several preprocessing steps were taken to account for as much unexplained variance in our data as possible (ie, we engineered features for 1- vs 2-handed typing mode, we parsed for within-user variation via random-effects, etc.), it is likely that certain sources of variation (eg, the specific digrams preferred by an individual, right-handed vs. left-handed, etc.) may still not be captured by our model, given that the data are naturalistically collected and we did not record the actual content typed.

Although our main findings are statistically significant, the effect sizes are likely small and thus may not be clinically ready to make inferences about differences between groups in real-world settings. Additional research and refinement will be required.

Last, cognition spans multiple recognized domains, yet it remains to be seen how our passive measurements may potentially map onto these distinct domains. Moreover, data collection in the wild is more susceptible to inattentiveness and distractions, potentially diluting the predictive strength of our metrics.

CONCLUSION

Our investigation over keystroke dynamics metadata derived from more than 14 million keypresses demonstrated the feasibility of using real-world passively collected keystroke dynamics to examine the relationship between typing performance, circadian rhythms, and depression symptom severity. Most importantly, we showed that more severe depression is associated with more variable typing speed, shorter keyboard sessions, and a decrease in typing accuracy. Taken as a whole, the main findings of our study support the promise of unobtrusively collected keystroke dynamics serving as temporally sensitive, digital biomarkers of mental health.

FUNDING

AUTHOR CONTRIBUTIONS

AL, SAL, AP, OA, PCN, and JD conceived the BiAffect project. AP, FH, EM, SY, DK, and KB developed the BiAffect iOS application. CV, HR, JZ, JS, JB, LO, RCM, CCB, APD, OA, and AL contributed to data collection and analysis, drafting of the article, and interpretation of the results. All authors pro-
vided critical feedback, assisted in the revision of the article, and approved of the final version.

SUPPLEMENTARY MATERIAL

ACKNOWLEDGMENTS

CONFLICT OF INTEREST STATEMENT

OA, RCM, and AL are co-founders of KeyWise AI. AL serves on the scientific advisory board of Buoy Health. OA serves on the board of Blueprint. JD is a consultant for Kitchry. DK is the president and founder of Arbormoon Software, Inc, and co-founded Gradient Valley with KB. CV, HR, JZ, JS, FH, AP, JB, SAL, SY, EM, LO, PCN, CCB, and APD report no biomedical financial interests or potential conflicts of interest.

REFERENCES

1. Howieson D. Current limitations of neuropsychological tests and assessment procedures. *Clin Neuropsychol* 2019; 33 (2): 200–8.
2. Riley E, Esterman M, Fortenbaugh FC, *et al*. Time-of-day variation in sustained attentional control. *Chronobiol Int* 2017; 34 (7): 993–1001.
3. Gamaldo AA, Allaire JC, Whitfield KE. Exploring the within-person coupling of sleep and cognition in older African Americans. *Psychol Aging* 2010; 25 (4): 851–7.
4. Chaytor N, Schmitter-Edgecombe M. The ecological validity of neuropsychological tests: A review of the literature on everyday cognitive skills. *Neuropsychol. Rev* 2003; 13 (4): 181–97
5. Seelye A, Hagler S, Mattek N, *et al*. Computer mouse movement patterns: A potential marker of mild cognitive impairment. *Alzheimers Dement (Amst)* 2015;1 (4): 472–80.
6. Trull TJ, Ebner-Priemer U. Ambulatory assessment. *Annu Rev Clin Psychol* 2013; 9 (1): 151–76.
7. Stange JP, Kleiman EM, Mermelstein RJ, *et al*. Using ambulatory assessment to measure dynamic risk processes in affective disorders. *J Affect Disord* 2019; 259: 325–36.
8. Buriro A, Akhtar Z, Crispo B, *et al*. Age, gender and operating-hand estimation on smart mobile devices. In: 2016 International Conference of the Biometrics Special Interest Group (BIOSIG); 2016.
9. Tsimperidis I, Rostami S, Katos V. Age detection through keystroke dynamics from user authentication failures. *Int J Digit Crime Forensics* 2017; 9: 1–16.
10. Pentel A. Predicting age and gender by keystroke dynamics and mouse patterns. In: UMAP '17: Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization. 2017:381–5.
11. Hilborn JV, Strauss E, Hultsch DF, *et al*. Intraindividual variability across cognitive domains: Investigation of dispersion levels and performance profiles in older adults. *J Clin Exp Neuropsychol* 2009; 31 (4): 412–24.
12. Kang JE, Lim MM, Bateman RJ, *et al*. Amyloid-β dynamics are regulated by orexin and the sleep-wake cycle. *Science* 2009; 326 (5955): 1005–7.
13. Chen R, Maljkovic V, Sunga M, *et al*. Developing measures of cognitive impairment in the real world from consumer-grade multimodal sensor streams. In: proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2019:2145–55.

14. Stringer G, Couth S, Brown LJE, *et al*. Can you detect early dementia from an email? A proof of principle study of daily computer use to detect cognitive and functional decline. *Int J Geriatr Psychiatry* 2018; 33 (7): 867–74.

15. Malhi GS, Kuiper S. Chronobiology of mood disorders. *Acta Psychiatr Scand* 2013; 128: 2–15.

16. McClung CA. Circadian genes, rhythms and the biology of mood disorders. *Pharmacol Ther* 2007; 114 (2): 222–32.

17. Cho CH, Lee HJ. Why do mania and suicide occur most often in the spring? *Psychiatry Investig* 2018; 15 (3): 232–4.

18. Moon JH, Cho CH, Son GH, *et al*. Advanced circadian phase in mania and delayed circadian phase in mixed mania and depression returned to normal after treatment of bipolar disorder. *EBioMedicine* 2016; 11: 285–95.

19. Krane-Gartiser K, Vaaler AE, Fasmer OB, *et al*. Variability of activity patterns across mood disorders and time of day. *BMC Psychiatry* 2017; 17 (1): 404. doi: 10.1186/s12888-017-1574-x.

20. Valenza G, Citi L, Lanatá A, *et al*. Revealing real-time emotional responses: a personalized assessment based on heartbeat dynamics. *Sci Rep* 2015; 4 (1): 4998.

21. Valenza G, Citi L, Gentili C, *et al*. Characterization of depressive states in bipolar patients using wearable textile technology and instantaneous heart rate variability assessment. *IEEE J Biomed Health Inform* 2015; 19 (1): 263–74.

22. Valenza G, Nardelli M, Lanatà A, *et al*. Wearable monitoring for mood recognition in bipolar disorder based on history-dependent long-term heart rate variability analysis. *IEEE J Biomed Health Inform* 2014; 18 (5): 1625–35.

23. Cho CH, Lee T, Kim MG, *et al*. Mood prediction of patients with mood disorders by machine learning using passive digital phenotypes based on the circadian rhythm: Prospective observational cohort study. *J Med Internet Res* 2019; 21 (4): e11029.

24. ResearchKit-Apple Developer. https://developer.apple.com/researchkit/ Accessed November 17, 2019.

25. Kroenke K, Strine TW, Spitzer RL, *et al*. The PHQ-8 as a measure of current depression in the general population. *J Affect Disord* 2009; 114 (1–3): 163–73.

26. Spitzer RL, Kroenke K, Williams J. Validation and utility of a self-report version of PRIME-MD: the PHQ Primary Care Study. *J Am Med Assoc* 1999; 282 (18): 1737–44.

27. Kroenke K, Spitzer RL, Williams J. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001; 16 (9): 606–13.

28. Oxman TE, Dietrich AJ, Williams JW, *et al*. A three-component model for reengineering systems for the treatment of depression in primary care. *Psychosomatics* 2002; 43 (6): 441–50.

29. Bellantuono C, Mazzi MA, Tansella M, *et al*. The identification of depression and the coverage of antidepressant drug prescriptions in Italian general practice. *J Affect Disord* 2002; 72 (1): 53–9.

30. McMillan D, Gilbody S, Richards D. Defining successful treatment outcome in depression using the PHQ-9: A comparison of methods. *J Affect Disord* 2010; 127 (1–3): 122–9.

31. Zuithoff NP, Vergouwe Y, King M, *et al*. The Patient Health Questionnaire-9 for detection of major depressive disorder in primary care: consequences of current thresholds in a crosssectional study. *BMC Fam Pract* 2010; 11 (1): 98.

32. Dawson EL, Caveney AF, Meyers KK, *et al*. Executive functioning at baseline prospectively predicts depression treatment response. *Prim Care Companion CNS Disord* 2017; 19(1): 16m01949.

33. Manczak EM, Skerrett KA, Gabriel LB, *et al*. Family support: a possible buffer against disruptive events for individuals with and without remitted depression. *J Fam Psychol* 2018; 32 (7): 926–35.

34. Zulueta J, Piscitello A, Rasic M, *et al*. Predicting mood disturbance severity with mobile phone keystroke metadata: a biaffect digital phenotyping study. *J Med Internet Res* 2018; 20 (7): e241–10.

35. Stange JP, Zulueta J, Langenecker SA, *et al*. Let your fingers do the talking: Passive typing instability predicts future mood outcomes. *Bipolar Disord* 2018; 20 (3): 285–8.

36. Singer JD, Willett JB. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford, United Kingdom: Oxford University Press; 2009.

37. Bates D, Mächler M, Bolker BM, *et al*. Fitting linear mixed-effects models using lme4. *J Stat Softw* 2015; 67 (1): 1–48.

38. West BT, Welch KB, Gałecki AT, *et al*. *Linear Mixed Models: A Practical Guide Using Statistical Software*. London: Chapman and Hall/CRC; 2014.

39. Kuznetsova A, Brockhoff PB, Christensen R. lmerTest Package: tests in linear mixed effects models. *J Stat Softw* 2017; 82 (13).

40. Mirman D. *Growth Curve Analysis and Visualization Using R*. London, United Kingdom: CRC Press; 2014.

41. Gallagher P, Nilsson J, Finkelmeyer A, *et al*. Neurocognitive intra-individual variability in mood disorders: Effects on attentional response time distributions. *Psychol Med* 2015; 45 (14): 2985–97.

42. Pinheiro JC, Bates DM. *Mixed-Effects Models in S and S-PLUS*. Berlin, Germany: Springer; 2000.