Multilingual markers of depression in remotely collected speech samples

Nicholas Cummins (nick.cummins@kcl.ac.uk)  
King's College London  https://orcid.org/0000-0002-1178-917X

Judith Dineley  
King's College London

Pauline Conde  
King's College London

Faith Matcham  
University of Sussex

Sara Siddi  
Universitat de Barcelona

Femke Lamers  
Amsterdam University Medical Centre, Vrije Universiteit

Ewan Carr  
King's College London

Grace Lavelle  
King's College London  https://orcid.org/0000-0003-3768-1797

Daniel Leightley  
King's College London  https://orcid.org/0000-0001-9512-752X

Katie White  
King's College London

Carolin Oetzmann  
King's College London  https://orcid.org/0000-0003-1480-917X

Edward Campbell  
University of Vigo

Sara Simblett  
King's College London  https://orcid.org/0000-0002-8075-8238

Stuart Bruce  
RADAR-CNS Patient Advisory Board

Josep Haro  
Parc Sanitari Sant Joan de Déu, Fundació Sant Joan de Déu

Brenda Penninx  
Amsterdam University Medical Centre, Vrije Universiteit

Yatharth Ranjan
Article

Keywords:

Posted Date: November 9th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2183980/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

Background

Speech contains neuromuscular, physiological, and cognitive components and so is a potential biomarker of mental disorders. Previous studies have indicated that speaking rate and pausing are associated with major depressive disorder (MDD). However, results are inclusive as many studies are small and underpowered and do not focus on clinical samples. These studies have also been unilingual and use speech collected in highly controlled settings. If speech markers are to help understand the onset and progress of MDD, we need to uncover markers that are robust to language and establish the strength of associations in real-world data.

Methods

We collected speech data in 585 participants with a history of MDD in the United Kingdom, Spain, and Netherlands as part of the RADAR-MDD study. Participants recorded their speech via smartphones every two weeks for 18 months. Linear mixed models were used to identify key cross-language markers of depression from a set of 28 speech features.

Findings:

Increased depressive symptoms were associated with speech rate, articulation rate and intensity of speech elicited from a scripted speech task. These three features had consistently stronger effect sizes than pauses.

Interpretation:

Participants with more severe depressive symptoms spoke more slowly and quietly, regardless of the language used. As decreases in speech rate and articulation rate had stronger effects than any of the pausing measures tested, we speculate the observed slowing of speech is most likely due to psychomotor impairments affecting articulation rates.

1. Introduction

Speech is uniquely placed in digital health: no other signal contains a combination of cognitive, neuromuscular, and physiological information. Speech is relatively simple to collect in daily life via remote measurement technologies (RMT) such as smartphones. Speech phenotypes could therefore become scalable digital biomarkers of health, providing insights into both current and predicted future health outcomes. A growing body of research has demonstrated associations between depression and
changes in specific acoustic and prosodic properties of speech.\textsuperscript{1,2} However, many of these findings are from small samples, are potentially underpowered, and should be treated as preliminary.

Most reported effects are observed in cross-sectional studies and include a range of prosodic and acoustic alterations such as flattened pitch contours and altered formant (measures of vocal tract resonances) trajectories (1,2). As well as being cross-sectional, most speech and depression research published in the last ten years has focused on model development using large multivariate feature spaces and machine learning paradigms, not on phenotype identification or speech feature characterisation (1,2). Most has also used only two publicly available datasets, the Audio-Visual Depressive Language (AViD) corpus and the Distress Analysis Interview Corpus (DAIC) (3–5). Both corpora are subject to two main limitations. Firstly, the speech collected cannot be regarded as clinical samples as they come from volunteers who have had their depression severity established through a questionnaire at the time of the study. Secondly, the metadata associated with both datasets is sparse, so potential confounding factors are unspecified.

Few observational speech-depression studies have assessed the predictive power of individual speech features. In a six-week study of 35 English-speaking participants, increases in pause time, and speaking rate, as well as decreases in the variation of second formant location were significantly associated with increasing depression severity (6). A subsequent, four-week study of 165 English-speaking participants observed that only increased speaking rate and pause time were associated with increased depression severity (7). Association between these measures and depression severity have been replicated. Participants of a clinical trial for treatment response in depression (N = 50) found that pauses became shorter and less variable as depression decreased (8). More recently, speech-to-pause ratio was associated with depression severity in a 4-week study of 18 English-speaking participants (9). Finally, a 10-week study of 241 Japanese-speaking participants also found increased speech rate and pauses were associated with depression (10).

These observational studies highlight the links between depression and changes in speaking rate and pausing (6–10). However, all studies were short-term, with the longest having only four observations over 21 weeks (8). They were also unilingual studies conducted on speech collected in highly controlled circumstances. No studies have examined how specific languages, a known source of variability in speech (11), affect these markers. Identifying language-independent markers of depression would increase the clinical effectiveness of speech-phenotypes, for example, by opening them up for inclusion in large multinational clinical trials. No studies have investigated speech parameters and depression over multiple time points with speech collected via RMT. Evaluating such associations in real-world data is vital to understanding the role that speech analysis could ultimately play in the management of chronic conditions such as depression.

We used data collected in the major European Innovative Medicines Initiative (IMI2) Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD) programme, a longitudinal cohort study examining the utility of multi-parametric RMT to predict changes in symptoms and relapse in
people with MDD (12). These data address previous limitations as the dataset: (i) contains speech samples from the largest clinical cohort study utilising RMT (13); (ii) collected longitudinally in the real world; and (iii) provides a multilingual dataset containing clinical samples.

We aimed to identify cross-language speech-based markers of depression from a smaller set of relevant features identified from the literature from remotely collected speech samples. We described the sociodemographic and clinical characteristics of the cohort and conducted analyses on these factors to determine potential determinants of data availability.

2. Methods

2.1 Study Design

RADAR-MDD was an observational cohort study of individuals with established MDD from three recruitment sites: King’s College London (KCL, London, United Kingdom); Amsterdam UMC, Vrije Universiteit (VUmc; Amsterdam, Netherlands); and Centro de Investigación Biomédica en Red del Área Salud Mental (CIBERSAM; Barcelona, Spain). The study protocol, eligibility and exclusion criteria have previously been reported (12,13). Briefly, the core eligibility criteria were having met the DSM-5 diagnostic criteria for non-psychotic MDD within the past two years prior to enrolment and having recurrent MDD (lifetime history of at least two episodes). All participants were aged over 18 and able to give written informed consent.

2.2 Ethics

Ethical approval

was obtained from the Camberwell St. Giles Research Ethics Committee (17/LO/1154) in London, from the Fundacio Sant Joan de Deu Clinical Research Ethics Committee (CI: PIC-128-17) in Barcelona, and from the Medische Ethische Toetsingscommissie VUmc (2018.012–NL63557.029.17) in Amsterdam.

2.3 Speech collection

RADAR-MDD was already an active study when speech collection began (13). Speech collection began in London in August 2019 and in Barcelona and Amsterdam in December 2019. Participants already enrolled at the start of speech collection were informed about speech collection in a newsletter and were provided with a link to a private YouTube instruction video. Participants who enrolled after the start of speech collection were briefed either face-to-face or remotely. The most important instructions were also provided in the purpose-made study smartphone application (app) (14).

Participants were asked to record themselves completing two speech elicitation tasks every two weeks. The recordings were collected via the RADAR-base active RMT (aRMT) data collection app. The app produced notifications each time speech recordings were scheduled. Before recording, participants were
reminded, via on-screen instructions, to find a quiet place to complete the recordings and speak in their normal voice.

The first activity was a *scripted speech task*, in which the participants read aloud an extract from Aesop's fable, *The North Wind and the Sun* (15); the extracts for each language are provided in Supplementary Tables 1–3. To minimise practice effects, the fable was split into three parts and participants were prompted to read a different extract at each recording. The second activity was a *free speech task*, in which participants were asked to speak about something they were looking forward to in the next seven days (6). In each task, participants were given the option of (i) re-recording their speech, for example, if they were interrupted while recording; and (ii) skipping the free-speech task.

As a safeguarding issue to discourage participants recording messages expressing suicidal ideation or intent, it was made clear to participants when they were introduced to the speech tasks that we would not be listening to the free speech audio while RADAR-MDD was an active study. Once recorded, the speech data were encrypted into a single file tagged with the participant's study ID number and sent to a secure server.

### 2.4 Speech data preparation

The collected data were decrypted into *Waveform Audio File Format* (WAV) files with a sampling frequency of 16 kHz and a 16-bit resolution; a separate file was created for each task. Some data could not be decrypted, so we define *audio files* as those files where we have a WAV file that can be played on standard audio editing software such as Audacity (16). Files that did not meet this criterion were not used in the analysis.

We then extracted a set of 28 speech features from the audio files. These features are partitioned into three groups: (1) Speech Timing Measures; (2) Prosodic; and (3) Phonation and Articulatory Measures. Features were extracted using Parselmouth (17), an open-source Python library that enables the use of Praat, a software package for speech analysis (18). Details on the features are provided in Supplementary Tables 4–6.

The following three criteria were used to determine if a file was included in our analysis. Firstly, files less than 2 seconds in length were removed from the analysis on the assumption that they were less likely to contain analysable speech. Secondly, an audio file was included if Parselmouth could return a value for all 28 features, otherwise we assume that there was a considerable amount of corrupting noise in the file. The final criterion was that a participant had to supply a minimum of two audio files for each task, i.e., the minimum number of files necessary for the data to be considered longitudinal.

### 2.5 Depression Assessments

We used the Inventory of Depressive Symptomatology – Self Report (IDS-SR) (19) and the 8-item Patient Health Questionnaire (PHQ-8) scale (20). The IDR-SR was used to identify the presence of *depressive symptoms* at baseline. For our analysis, we define baseline depression to be the IDS-SR score obtained
within six weeks of a participant first being scheduled to participate in the speech task. The PHQ-8, which gives a self-reported depressive symptom severity, was collected remotely and concurrently with the speech recordings, every two weeks, via the aRMT RADAR-Base app (14). Given speech samples and PHQ-8 were collected concurrently, we use this measure when identifying key cross-language speech-based markers of depression.

2.6 Patient Involvement

The RADAR-MDD protocol was co-developed with a patient advisory board who shared their opinions on several user-facing aspects of the study including the choice and frequency of survey measures, the usability of the study app, participant-facing documents, selection of optimal participation incentives, selection, and deployment of wearable device as well as the data analysis plan. The speech task, and subsequent analysis has been discussed specifically with the RADAR-CNS Patient Advisory Board (PAB), and a member of PAB is also a co-author of this manuscript.

3. Analyses

3.1 Sociodemographic and clinical determinants of speech data availability

We conducted analyses to identify differences in the availability of speech data based on sociodemographic and clinical variables likely to predict missingness and included depression severity (IDS-SR score), age, gender, height (as a proxy of vocal tract length), and years of education (proxy for reading ability). Age, height, gender, and years spent in education are included in our analysis as these are long-term speaker-traits which can affect acoustic and prosodic speech markers (21,22).

We first described each measure using medians and interquartile range and then used Chi-squared (gender) and Wilcoxon signed-rank (age, height, years of education, depression severity) tests to determine whether there were differences in the proportion of participants providing analysable speech samples. This analysis was conducted using IBM SPSS Statistics software, and was performed separately by language (English, Dutch, Spanish) and for the ‘scripted’ and ‘free speech’ tasks in turn.

3.2 Relationship of speech markers with depression

We used linear mixed effect models (LMEs) to estimate associations between fortnightly PHQ-8 depression scores and 28 speech features measured over the preceding two weeks. The availability of speech data varies across our cohort (13); LMEs allowed us to include all available information from each participant (23,24). We included an individual random intercept to account for the clustering of repeated fortnightly assessments within each participant. Each speech feature was standardised (mean = 0; standard deviation = 1) to improve estimation and interpretability. Age, height, gender, and years spent in education are included as. Again, all models were estimated separately by language and by tasks.
Models were estimated using the lme4 package for R (25). We reported standardised coefficients and 95% bootstrap confidence intervals (1000 bootstrap iterations).

**4. Results**

**4.1 Cohort characteristics**

A total of 585 participants were enrolled in RADAR-MDD during the speech collection period. The largest cohort was the United Kingdom with 325 participants (56%), followed by Spain with 143 participants (24%), and then the Netherlands with 117 participants (20%). All cohorts have a larger female representation; 78% in the Netherlands, 76% in the United Kingdom, and 71% in Spain. The depression scores at baseline indicate in each country fall in the moderate severity range for the IDS-SR. Full details of the distribution of our sociodemographic and clinical variables are given in Supplementary Table 7.

The final analytical sample contained 461 (79%) individuals who had analysable data on one or both tasks (457/585 (78%) ‘scripted’ task, 435/585 (74%) ‘free response’ task, 431/585 (74%) provided information for both tasks – see Fig. 1). No baseline demographic and clinical depression characteristics were associated with who did or did not provide analysable speech data in either speech tasks for the British (Supplementary Table 8) and Dutch (Supplementary Table 9) cohorts. Baseline depression severity was significantly higher (p = .024) for the Spanish participants who provided analysable scripted speech, while Years in Education was significantly higher (p = .009) for the Spanish participants who provided analysable free-response speech (Supplementary Table 10).

**Figure 1**

**4.2 Provision of longitudinal speech samples**

Speech collection was active in RADAR-MDD for a period of 620 days. The median speech collection period was 433 days (interquartile range (IQR): 358–473 days, range: 4-590 days). As speech recording was scheduled once every two weeks, this resulted in a median of 31 recording opportunities (interquartile range (IQR): 26–34, range: 1–43). A more detailed breakdown of the number of file analysed for the scripted and free response task for each country, as well as descriptive statistics of the corresponding PHQ-8 files are given in Supplementary Table 11. A comparison of the PhQ-8 distributions (per language, per speech task) is given in Supplementary Fig. 1.

**4.3 Associations between speech features and depression severity**

**Speech timing**

For the scripted task, we found speaking and articulation rates to be negatively associated with depressive severity in all three languages (Fig. 2; Supplementary Table 12) indicating that participants with more severe depressive symptoms spoke more slowly, regardless of the language being spoken. For
example, a 1 SD increase in speaking rate was associated with a 0.2 unit (95% CI), 0.44 unit (95% CI) and a 0.27 unit (95% CI) decrease in the subsequent PHQ-8 score, in the UK, Dutch and Spanish cohorts, respectively.

Recording duration was positively associated with depressive severity in the UK and Dutch cohorts, and there was a no-significant trend in the Spanish data. Phonation ratio (positive) and number of pauses (negative) were associated with depressive severity for the UK cohort only, with similar no-significant trends observable in the Dutch and Spanish cohorts.

For the free response task, phonation ratio, speaking rate and articulation rate were negatively associated with depressive severity in the UK and Dutch cohorts, while average syllable duration was positively associated (Fig. 3; Supplementary Table 12). A larger negative $\beta$ coefficient indicates a similar trend for phonation ratio in the Spanish data but no evidence of associations in the other features. Recording duration (negative), phonation time (negative), number of syllables (negative), mean length run (negative), and average pause duration (positive) were associated with depressive severity for the UK cohort only. Similar trends can be seen for phonation time (Dutch, Spanish), number of syllables (Dutch, Spanish), mean length run (Dutch), and average pause duration (Dutch) in the other language groups, but not at a significant level.

Prosodic and Phonation

In the scripted task, as with speaking rate and articulation rate, we found that intensity was negatively associated with depressive severity in all three languages (Fig. 2; Supplementary Table 13). A 1 SD increase in speaking rate was associated with a 0.28 unit (95% CI), 0.43 unit (95% CI) and a 0.34 unit (95% CI) decrease in the subsequent PHQ-8 score, in the UK, Dutch and Spanish cohorts, respectively. Harmonic to noise ratio (HNR) was positively associated with depressive severity in the UK and Dutch cohorts, while shimmer was negatively associated. A similar, but not significant trend can be seen for jitter in the Spanish cohort. Finally, mean pitch was negatively associated with depressive severity for the UK cohort only, with a non-significant negative trend in the Dutch cohort.

For the free response task, mean pitch (negative), intensity (negative), and fraction of unvoiced frames (positive) were associated with depressive severity in the UK and Dutch cohorts (Fig. 3; Supplementary Table 13). Mean pitch and intensity also display negative tendencies in the Spanish cohort. Jitter was positivity associated with depressive severity in the UK and Spanish cohorts. The number of voice breaks (UK) and pitch standard deviation (Dutch) were negatively associated with depression.

Articulatory Measures

Associations between articulatory features and depression severity were observed within single languages only. Each country returned one associated formant change in the scripted task (Fig. 2; Supplementary Table 14). Decreasing mean F1 frequency, decreasing standard deviation in F2 bandwidth, and increasing standard deviation in F2 frequency with increasing depression were observed...
in the UK, Dutch, and Spanish data respectively. There is evidence of decreasing mean F1 frequency in the Spanish data, however not at a significant level. The only association observed in the free response task was an increase in mean F1 bandwidth with increasing depression in the Dutch cohort (Fig. 3; Supplementary Table 14).

4. Discussion

The RADAR-MDD speech dataset is unique in its scale, duration and the number of languages recorded. Participants provided speech samples for around 62 weeks, compared to a maximum 21 weeks in previous studies (8). The dataset also provides speech recordings in three languages, all collected using the same speech elicitation tasks and software platform in contrast to other datasets that contain only one language (1,2). We found that as depression increases, participants spoke slower and more quietly, whatever language they used. Decrease in speaking rate have been observed in previous longitudinal studies (6–10), this is the first time that decrease in intensity with increasing levels of depression have been observed in a longitudinal study.

There are two ways to slow speech rate; the insertion of longer pauses, or decreasing the rate of speech sound production (1,26). The insertion of pauses is linked with cognitive impairments, while the decrease in the rate of speech sound production is more reflective of psychomotor impairments (26). As decreases in speech rate and articulation rate have stronger effects than any of the pausing measures, we can infer that decreases in speech rate are due to increases in phonation time rather than increases in pause rate. The changes we observed are therefore more likely to be due to increases in neuromuscular impairment affecting the rate of speech production. Decreases in intensity with increased levels of depression are not universally reported in the literature (1,2). Most studies which do not report a significant association are based on small samples and so may be underpowered.

Out of the pausing measures tested, number of pauses (scripted) and average pause duration (free response) were significantly associated with depression in the UK cohort. Most other studies that have reported this finding were conducted in English (6–9), so it is conceivable that it is a more English specific effect. We found no evidence of associations between pause rate and depressive symptoms in any speech tasks or language, in contrast to previous findings.6–10 This discrepancy may be due to the statistics (we took account of clustering of repeated measures within individuals) and none of the other studies were conducted in real world settings, meaning the speech collected could also be subject to the observer paradox (27). This is a phenomenon in which the speaking style of a participant changes due to the presence of a researcher or clinician during the recording session. While we did observe a small number of format changes associated with depression severity, there were no consistent cross-language results for these features. Studies that have previously reported significant formant findings have either used different speech elicitation methods, i.e., extended vowel sounds (6), or were cross-sectional (28).

We have identified four main limitations of this study. These findings are across groups and so have limited impact on identifying changes within an individual which is the focus of our planned future
research. Secondly, we did not collect the data in a laboratory so our observed intensity effects could be artefacts in participant interactions with the recording equipment. However, given the number of recordings collected, this is unlikely. Thirdly, we analysed features extracted over the entire duration of the recordings and a phonetic transcription of the data would enable a more fine-grained analysis of different voice quality and formant effects so we may have underestimated their importance. Finally, despite the long data collection period, 125 participants (21%) did not provide longitudinal data for either speech task. No clinical or demographic characteristic predicted whether people took part in the speech tasks which is important as it suggests that depression severity is unlikely to influence the willingness to participate in speech tasks. But participation varied with education level in the Spanish free-speech task, highlighting the need to understand facilitators and barriers of remote speech collection (29,30). Future work will include in depth analysis to understand if there are specific people who are less willing to provide speech data.

To conclude, this study highlights that the collection of long-term in-the-wild longitudinal speech data from a MDD cohort is feasible. As our findings are based on multilingual data they represent a considerable step-change in the usefulness of speech as a digital phenotype of MDD. Importantly, as the identified features are from the scripted task, they are potentially more favourable to participants in future studies from a privacy perspective. Combining the results of this study with previously presented analysis (6–10), there is strong evidence to support the use of speech-rate measures as digital phenotypes of MDD in larger scale research projects. Future work, including the use of these features in multivariate prediction models will increase the strength of associations between other speech severity and depression severity.

Declarations

Funding: This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 115902. This joint undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

Data Availability:

Due to the confidential nature of speech data, we are unable to this data publicly available. Access to the data used for can be made through reasonable requests to the RADAR-CNS consortium and will be subject to local ethics clearances. Please email the corresponding author for details.

Competing Interests

The Authors declare no Competing Financial or Non-Financial Interests.

Acknowledgement:

The RADAR-CNS project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 115902. This Joint Undertaking receives support from the
European Union’s Horizon 2020 research and innovation programme and EFPIA (www.imi.europa.eu). This communication reflects the views of the RADAR-CNS consortium and neither IMI nor the European Union and EFPIA are liable for any use that may be made of the information contained herein. The funding body have not been involved in the design of the study, the collection or analysis of data, or the interpretation of data.

Participant recruitment in Amsterdam was partially accomplished through Hersenonderzoek.nl, a Dutch online registry that facilitates participant recruitment for neuroscience studies (https://hersenonderzoek.nl/). Hersenonderzoek.nl is funded by ZonMw-Memorabel (project no 73305095003), a project in the context of the Dutch Deltaplan Dementie, Gieskes-Strijbis Foundation, the Alzheimer’s Society in the Netherlands and Brain Foundation Netherlands Participants in Spain were recruited through the following institutions: Parc Sanitari Sant Joan de Déu network of mental health services (Barcelona); Institut Català de la Salut primary care services (Barcelona); Institut Pere Mata-Mental Health Care (Tarrassa); Hospital Clínico San Carlos (Madrid). This paper represents independent research part funded by the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King’s College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

We thank all the members of the RADAR-CNS patient advisory board for their contribution to the device selection procedures, and their invaluable advice throughout the study protocol design. This research was reviewed by a team with experience of mental health problems and their carers who have been specially trained to advise on research proposals and documentation through the Feasibility and Acceptability Support Team for Researchers (FAST-R): a free, confidential service in England provided by the National Institute for Health Research Maudsley Biomedical Research Centre via King’s College London and South London and Maudsley NHS Foundation Trust.

We thank all GLAD Study volunteers for their participation, and gratefully acknowledge the NIHR BioResource, NIHR BioResource centres, NHS Trusts and staff for their contribution. We also acknowledge NIHR BRC, King’s College London, South London and Maudsley NHS Trust and King’s Health Partners. We thank the National Institute for Health Research, NHS Blood and Transplant, and Health Data Research UK as part of the Digital Innovation Hub Programme.

We thank our colleagues both within the RADAR-CNS consortium and across all involved institutions for their contribution to the development of this protocol. We thank all the members of the RADAR-CNS patient advisory board for their contribution to the device selection procedures, and their invaluable advice throughout the study protocol design.

Author Contributions

NC: Conceptualization, Methodology, Formal Analysis, Writing Original Draft; JD: Conceptualization, Methodology, Data Curation, Writing – Original Draft; PC: Data Curation, Software; FM: Conceptualization,
References


Figures

Figure 1

Breakdown of number of participants supplying longitudinal speech data in RADAR-MDD. *Never Attempt* denotes the number participants who never attempted that task; *Corrupt Files Only*, are participants who attempted the task, but created only corrupt files; *Attempted Once*, are participants provided only one file; and *No Analysable Data* are participants excluded as their files failed the paper’s inclusion of being both over 2 seconds in length and returned usable Parselmouth features.
Figure 2. Association of speech features with PHQ-8 score for "Scripted" tasks (n=457 individuals; 7356 observations)

Points represent the difference in PHQ-8 per 1 SD difference in each feature. 95% confidence intervals obtained using bootstrap resampling (1000 samples).

See image above for figure legend.
Figure 3. Association of speech features with PHQ-8 score for "Free response" tasks (n=435 individuals; 6106 observations)
Points represent the difference in PHQ-8 per 1 SD difference in each feature. 95% confidence intervals obtained using bootstrap resampling (1000 samples).

See image above for figure legend.

Supplementary Files
This is a list of supplementary files associated with this preprint. Click to download.
• SupplementaryMaterial.docx