npj | Schizophrenia

www.nature.com/npjschz

PERSPECTIVE OPEN

More than a biomarker: could language be a biosocial marker of psychosis?

Lena Palaniyappan[1,2,3,4]✉

Automated extraction of quantitative linguistic features has the potential to predict objectively the onset and progression of psychosis. These linguistic variables are often considered to be biomarkers, with a large emphasis placed on the pathological aberrations in the biological processes that underwrite the faculty of language in psychosis. This perspective offers a reminder that human language is primarily a social device that is biologically implemented. As such, linguistic aberrations in patients with psychosis reflect both social and biological processes affecting an individual. Failure to consider the sociolinguistic aspects of NLP measures will limit their usefulness as digital tools in clinical settings. In the context of psychosis, considering language as a biosocial marker could lead to less biased and more accessible tools for patient-specific predictions in the clinic.

npj Schizophrenia (2021)7:42 ; https://doi.org/10.1038/s41537-021-00172-1

> "I have resisted the term sociolinguistics for many years, since it implies that there can be a successful linguistic theory or practice which is not social"[1]

COMPUTATIONAL LINGUISTICS IN PSYCHOSIS

In recent times, a surge of methodological advances in sampling human discourse has brought the spotlight back on the phenomenon variously known as 'formal thought disorder' (FTD), 'speech disorder' or 'communication disturbances' in schizophrenia[2–5]. The emphasis on objective measurement of speech in real-world settings using automated assessment is not new[6]. But the improved access to normative corpora for analysis, the corpus-independent graph metrics and the emergence of ambulatory approaches for speech capture that aid Artificial Intelligence based learning systems have improved scalability, and provided the much needed momentum to this field of inquiry. Several predictive as well as mechanistic studies have been published in npj Schizophrenia[7–11] in recent times, with excellent state-of-art appraisals of computational linguistics found elsewhere[2,4,5,12]. The assessment of thought disorder has extended from being a qualitative assessment using FTD scales to a more quantitative determination based on automated measures.

To date, the emphasis of quantitative speech studies has been a utilitarian one. This involves examining language as a biomarker of pathogenic neural processes of schizophrenia, and leveraging this biomarker for diagnostic and prognostic purposes. In fact, natural language processing (NLP) measures, often digitally acquired, have been evaluated as biomarkers of risk[7,13–15], diagnosis/prognosis[8,16], and to study pharmacological effects[10] in recent times. In this broad context of use, NLP measures are treated as many other biomarkers investigated in psychiatry[17].

SOCIOMARKERS, BIOMARKERS AND BIOSOCIAL MARKERS

The term 'sociomarkers' is sometimes used for objective characteristics that mark a social condition or process that an individual has experienced or currently embedded in. For example, neighbourhood housing quality could be a social marker to predict hospital visits in children with asthma[18]. Some sociomarkers (e.g., homelessness) can also be used as a proxy endpoint for systems-interventions in multidisciplinary mental health settings. However, variations in these systems-level sociomarkers (e.g., homelessness) are often not attributable to changes in human biology. In contrast, quantity and quality of speech (or written language) can be manipulated via several biological interventions (e.g., ketamine infusion[19], neuromodulation[20], vascular insults etc. to name a few). These linguistic measures are not only objectively quantifiable but also change with disease processes; thus, linguistic markers satisfy the broadly accepted criteria for a biomarker[21].

While several biomarkers are affected by social factors, their quantification process itself is unlikely to be socially influenced. The structure of human language is heavily influenced by sociocultural[22] and contextual factors[23,24], in ways that are much deeper than the effect of these factors on conventional tissue-based biomarkers. First, our everyday speech is replete with markers of our present and past social states[25] e.g., schooling, level of one's education, current social network to name a few. What we say, and how we say it, depends on the immediate social context in which a speech event is embedded. The social and professional status of the speakers, in-group vs. out-group differences, class, ethnicity, age, gender as well as the social relatedness[23] provide this social context. Quantitative linguistic markers such as lexical diversity are strongly correlated with

[1]Department of Psychiatry, Schulich School of Medicine and Dentistry, University of Western Ontario, London, ON, Canada. [2]Department of Medical Biophysics, Schulich School of Medicine and Dentistry, University of Western Ontario, London, ON, Canada. [3]Robarts Research Institute, University of Western Ontario, London, ON, Canada. [4]Lawson Health Research Institute, London, ON, Canada. ✉email: lpalaniy@uwo.ca

socioeconomic markers such as neighbourhood income levels and density[26]. More relevant to the study of psychosis, is the large body of literature supporting the effect of early life socioeconomic status on quantitative speech markers (often referred to as the "million word-gap"[27,28]). Besides lexical diversity, social determinants such as social class affect syntactic complexity[29], while parental education status relates to frequency of certain parts of speech in adult speakers. For example, patients with schizophrenia born to better educated parents use more conjunction, less personal and first person singular pronouns[30]. Graph-based linguistic markers, considered to be predictive of diagnostic outcomes in psychosis[8,15], relate to family income and more strongly, to the number of years of school education[31]. Features of formal thought disorder are pronounced in maltreated children[32]; in terms of quantitative markers, the experience of institutional care has profound effects on lexical diversity and mean length of utterances during childhood, as demonstrated by the remarkable randomized trial of foster care vs. institutional care in Bucharest[33,34]. Such quantitative differences in early lexical development can affect later acquisition of syntax, and the processing dynamics in adult speakers[35] (also see[36]).

Linguistic structure that contributes to the various quantitative speech markers used in the study of psychosis, varies with the language being studied. Parts of speech that are reportedly abnormal in schizophrenia, are not universally present across languages. For example, in psychosis, formal thought disorder in English speakers relates to excess use of pronouns[37,38] and discourse connectives[39], but reduced use of articles and prepositions relative to other function words[40,41]. In particular, first person singular pronouns (which by definition excludes referring to the listener) are increased[38,42] while first person plural pronouns (which may or may not exclude the listener) are reduced in the presence of psychosis[36]. Such a pattern cannot be gleaned if we examine patient speech transcripts in a different language, Tamil, for example. Tamil, one of the few living classical languages, splits the first person plural to inclusive and exclusive versions, second person pronouns to singular and plural versions, has no articles, uses suffixes rather than unbounded discourse connectives, and has no prepositions but only postpositions[43,44]. Bilingual speakers whose first language (L1) structure varies from English (L2), often continue to make subtle structural errors[45,46] and display altered semantic coherence[47,48]. Even within a given language, the performance of NLP algorithms can vary with the dialect being spoken[49]. This increases the risk of some speakers being misclassified as being deviant from other healthy subjects, if such speech markers are used without due consideration of sociolinguistic differences. Thus, speech markers of psychosis derived from one language, or one dialect, may not always perform well for a patient speaking a different dialect, or has a different L1. This introduces another layer of socio-developmental variability with no biological causal basis.

In summary, the early life language environment is likely to have a critical effect on several quantitative speech markers measured in adult speakers. Linguistic measures are not merely markers of a biological state; they are also 'fossils' of one's social circumstances. Thus, language is best considered as a biosocial marker. To this end, a significant portion of inter-individual variance in linguistic structure among patients with psychosis could also relate to social factors. This speculation, based on the data from non-psychotic individuals reviewed above, is indirectly supported by several empirical observations linking qualitative measures of thought disorder with social factors in psychosis.

SOCIAL DETERMINANTS OF 'FORMAL THOUGHT DISORDER'

Language difficulties appear long before a clinical or preclinical state can be defined in psychosis[50,51]. Specific qualitative features of formal thought disorder that are apparent after the onset of psychosis relate to various social factors such as social class, educational exposure, immigrant status and social isolation in patient samples. When studying the generalisability of Thought Disorder Index (TDI), Haimo and Holzman observed that lower social class related to higher TD scores in healthy subjects, but to lower TD scores among patients with schizophrenia[52], highlighting the differential role that social class can play in the assessment of TD. Poverty of thought is more pronounced in the less educated patients[39], and those with parents from lower socioeconomic status (especially in female patients[53]). More recently, Nogueira and colleagues reported a relative excess of formal thought disorder among individuals with familial migration history at ultra-high risk of psychosis, though this study was limited by its sample size[54]. Nevertheless, these findings resonate with Berg and colleagues[55] who showed in a large Norwegian sample ($n = 1081$) that while natives and immigrants had mostly similar symptom profile, the most prominent difference in symptom dimensions, especially in visible minorities, related to disorganisation factor of PANSS (especially, difficulties in abstract thinking). Social isolation is associated with disorganised thinking more than any other positive symptoms of psychosis[56,57]. A reduction in exposure to social dialogue and feedback due to extended social isolation that occurs in the face of immigration, frequent school dropouts[58], institutionalisation may affect the ability of perspective-taking, increasing the risk of speech disturbances[59].

While some of the above associations can result from a reverse causality e.g., poverty of thought contributing to educational failure and disorganisation contributing to social isolation, other social factors precede the onset of language disturbances and cannot be assumed to result from a patient's TD per se (e.g., parental immigration, parental socioeconomic status, minority social status). Taken together, a broad range of social states affect the degree of language disturbances in psychosis, though most of the existing evidence comes from examinations of qualitative deviations (i.e., formal thought disorders). Insofar as the quantitative NLP markers relate to TD, they may also be altered by the several interacting social factors that influence TD; this requires systematic evaluation in future studies.

IS LANGUAGE THE ONLY MARKER TO BE SOCIALLY INFLUENCED?

Given the promising translational value of computational linguistic markers in the era of digital health, this Perspective focuses specifically on language. Nevertheless, many putative biomarkers are likely influenced by social factors; one may argue that linguistic markers are not exceptional in this regard. For example, social adversities affect endocrine, metabolic and neural biomarkers[60], though some markers are affected more by social factors than the other[61,62]. This raises the question whether linguistic markers of disease states are more 'socioplastic' than non-linguistic biomarkers. At present, we do not know if social factors affect linguistic markers more than they affect the other putative biomarkers of psychosis (e.g., hippocampal volume, cortical gyrification). We also do not know if social factors explain proportionately more variance than biological factors for putative linguistic markers in psychosis. Given the lack of data from clinical samples, it is worth considering certain factors that indirectly imply the relative importance of social factors on linguistic markers.

Linguistic markers (both verbal and nonverbal) carry explicit information that can identify individuals to the social groups or places to which they belong[63]. Some aspects of language can even be altered intentionally in response to one's social needs (e.g., code-switching and thus the choice of words and syntax[64]). In addition, linguistic markers can also be affected by the context in which the measurement is undertaken (e.g., the formality and the familiarity of the receiver affects markers of linguistic sentiment[65]), supporting

Table 1. Social factors with reported influence on quantitative speech markers.

Speech markers relevant to psychosis	Non-biological (social) factors with demonstrable influence in healthy subjects
Graph connectivity	Family income Years of education
Length of utterance	Institutionalization
Lexical diversity	Parental socioeconomic status Neighbourhood Institutionalization
Parts of speech (function words)	Parental education Bilingualism Dialect variations
Pauses and prosody	Binary entity described as race[a] Dialect variations
Semantic coherence	Binary entity described as race[b]
Semantic density	Bilingualism
Syntactic complexity	Social class

All listed speech markers are identified as promising candidates for computational linguistic analysis in psychosis (see Hitczenko and colleagues[12]). Individual studies with supporting evidence are discussed in the text. Most of the included studies were not specifically powered to examine the reported associations; the associations listed in this table are best considered as signals that require further systematic evaluation and not as conclusive links.

[a] Race was evaluated as a binary variable (Black or African-American/White)[66,67], without a multidimensional assessment of underlying components[75].

the argument that participant-level speech markers "on closer examination prove to be markers not of participant per se, but of participant in a particular situation"[23]. These social contributors (*identity*, *necessity*, *familiarity* and *situationality*) are unique to language and cannot be significant influences on illness-related biomarkers such as brain volume or gyrification.

In the wake of the circumstantial evidence reviewed here, the influence of social factors on linguistic markers of psychosis requires further scrutiny. Irrespective of the eventual influence apportioned to biological vs. social factors, neglecting one aspect in favour of the other risks oversimplification and may be detrimental to the development of computational linguistic applications in psychosis.

IMPLICATIONS AND FUTURE DIRECTIONS

Speech is an easily accessible marker to monitor psychosis; with NLP approaches, the highly desired objectivity in speech analysis now appears achievable. But this objectivity does not confer a complete freedom from social influences to NLP measures of language in psychosis. For example, Hitczenko and colleagues recently observed that sociodemographic differences in certain NLP measures of coherence are larger than the differences related to psychosis at-risk state[66], though this study only included a small number of subjects. Such outcomes may be the result of inherent biases in the corpora used for word embedding models as discussed elsewhere[67,68]; at present, we do not know the degree to which such socially driven biases contribute to prediction accuracies of NLP algorithms in clinical cohorts with psychosis. The relative influence of social factors is unlikely to be uniform across the various quantitative speech measures. NLP based predictive studies in psychosis need to urgently consider the influence of factors such as parental/individual socioeconomic status, neighbourhood deprivation levels, education or immigration status on the various structural speech markers that are under scrutiny[12]. (See Table 1)

As specific applications of NLP markers for psychosis become more clear, rigorous tests against evidentiary criteria[69] will be required to enable clinical use. At this stage, the performance of NLP assays (i.e., predictive consistency, dose-response relationships, generalizability) and their clinical interpretability will depend on the effects of social factors on these assays. For example, if transition to psychosis is more prevalent among individuals that dropped out of school, and if such individuals also have lower semantic coherence (a NLP measure), the predictive performance of semantic coherence for transition may only reflect the contribution of poor school retention. When applied to a social milieu with uniformly better school retention, the algorithm will have a poor yield, as the variance in semantic coherence will be now reduced. This does not necessarily diminish the mechanistic importance of the language disturbance per se, as one can conceive a causal link between linguistic issues and school dropout, and both being predictive of psychosis. Nevertheless, the relative utility of this NLP measure in prognostic prediction is likely to be reduced if its relationship with readily measurable and differentially distributed social determinants is not fully considered. Besides analytical and clinical validations[70], the widespread use of digital language-based biomarkers in psychosis requires a careful evaluation of the social determinants of linguistic aberrations highlighted here.

The view of NLP outputs as biosocial markers rather than biomarkers has two important implications. First, this reminds us that the algorithmic bias, defined as the amplification of existing social inequities when employing Artificial Intelligence algorithms[71], is a critical issue when using linguistic data. Second, the biosocial emphasis highlights the need for social diversity in the participant recruitment to obtain meaningful predictive values[72]. If unattended, these issues may ordain the NLP based algorithms to a limited utility (poor incremental value). In an extreme case scenario, such NLP-based predictions may paradoxically be less useful for the underrepresented groups, for whom prognostic prediction is likely to be most valuable.

An ideal NLP-based algorithm should operate in an unbiased manner, have equitable performance across different health systems, and provide actionable results. We require certain interventions to achieve this goal. NLP studies in psychosis must seek diverse samples and report social indicators with diligent detail. Besides predictive accuracy, algorithms must be examined for contextual specificity e.g., test performance at different social class or immigration strata. Generalisability across social groups/contexts should not be automatically assumed before such efforts are undertaken. With context-specific performance metrics, we will be able to build clinically meaningful counterfactual explanations[73] for NLP-based test results. Third, attempts to quantify the degree of social influence on the various quantitative linguistic markers is required to build multilevel models without redundancy. Building a large corpus of multilingual 'benchmark dataset'[74] for psychosis that can capture sufficient ethnic, cultural, social, economic, educational and lifestyle differences is a crucial step in this regard.

Received: 30 March 2021; Accepted: 6 August 2021;
Published online: 31 August 2021

REFERENCES

1. Labov, W. Linguistics and Sociolinguistics. in *Sociolinguistics: A Reader* (eds. Coupland, N. & Jaworski, A.) 23–24 (Macmillan Education UK, 1997). https://doi.org/10.1007/978-1-349-25582-5_3.
2. Corcoran, C. M. & Cecchi, G. Using language processing and speech analysis for the identification of psychosis and other disorders. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* (2020) https://doi.org/10.1016/j.bpsc.2020.06.004.

3. Minor, K. S., Willits, J. A., Marggraf, M. P., Jones, M. N. & Lysaker, P. H. Measuring disorganized speech in schizophrenia: automated analysis explains variance in cognitive deficits beyond clinician-rated scales. *Psychol. Med.* **49**, 440–448 (2019).
4. de Boer, J. N., Brederoo, S. G., Voppel, A. E. & Sommer, I. E. C. Anomalies in language as a biomarker for schizophrenia. *Curr. Opin. Psychiatry* **33**, 212–218 (2020).
5. Low, D. M., Bentley, K. H. & Ghosh, S. S. Automated assessment of psychiatric disorders using speech: a systematic review. *Laryngoscope Investig. Otolaryngol.* **5**, 96–116 (2020).
6. Elvevåg, B., Foltz, P. W., Weinberger, D. R. & Goldberg, T. E. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophr. Res.* **93**, 304–316 (2007).
7. Bedi, G. et al. Automated analysis of free speech predicts psychosis onset in high-risk youths. *Npj Schizophr.* **1**, 1–7 (2015).
8. Mota, N. B., Copelli, M. & Ribeiro, S. Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. *Npj Schizophr.* **3**, 1–10 (2017).
9. Rezaii, N., Walker, E. & Wolff, P. A machine learning approach to predicting psychosis using semantic density and latent content analysis. *Npj Schizophr.* **5**, 1–12 (2019).
10. de Boer, J. N., Voppel, A. E., Brederoo, S. G., Wijnen, F. N. K. & Sommer, I. E. C. Language disturbances in schizophrenia: the relation with antipsychotic medication. *Npj Schizophr.* **6**, 1–9 (2020).
11. Stanislawski, E. R. et al. Negative symptoms and speech pauses in youths at clinical high risk for psychosis. *Npj Schizophr.* **7**, 1–3 (2021).
12. Hitczenko, K., Mittal, V. A. & Goldrick, M. Understanding language abnormalities and associated clinical markers in psychosis: the promise of computational methods. *Schizophr. Bull.* **47**, 344–362 (2021).
13. Gutiérrez, E. D., Cecchi, G., Corcoran, C. & Corlett, P. Using automated metaphor identification to aid in detection and prediction of first-episode Schizophrenia. in *Proceedings of the 2017 conference on empirical methods in natural language processing* 2923–2930 (Association for Computational Linguistics, 2017). https://doi.org/10.18653/v1/D17-1316.
14. Corcoran, C. M. et al. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry* **17**, 67–75 (2018).
15. Spencer, T. J. et al. Lower speech connectedness linked to incidence of psychosis in people at clinical high risk. *Schizophr. Res.* **228**, 493–501 (2021).
16. Palaniyappan, L. et al. Speech structure links the neural and socio-behavioural correlates of psychotic disorders. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **88**, 112–120 (2019).
17. García-Gutiérrez, M. S. et al. Biomarkers in psychiatry: concept, definition, types and relevance to the clinical reality. *Front Psychiatry.* **11**, 432 (2020).
18. Shin, E. K., Mahajan, R., Akbilgic, O. & Shaban-Nejad, A. Sociomarkers and biomarkers: predictive modeling in identifying pediatric asthma patients at risk of hospital revisits. *Npj Digit. Med.* **1**, 1–5 (2018).
19. Nagels, A. et al. S -Ketamine-induced NMDA receptor blockade during natural speech production and its implications for formal thought disorder in Schizophrenia: a pharmaco-fMRI study. *Neuropsychopharmacology* **43**, 1324–1333 (2018).
20. Murakami, T., Ugawa, Y. & Ziemann, U. Utility of TMS to understand the neurobiology of speech. *Front Psychology.* **4**, 446 (2013).
21. Robin, J. et al. Evaluation of speech-based digital biomarkers: review and recommendations. *Digit. Biomark.* **4**, 99–108 (2020).
22. Beckner, C. et al. Language is a complex adaptive system: position paper. *Lang. Learn.* **59**, 1–26 (2009).
23. Brown, P. & Fraser, C. Speech as a marker of situation. in *Social markers in speech* 33–62 (Cambridge University Press, 1979).
24. Roberts, G. Perspectives on language as a source of social markers. *Lang. Linguist. Compass* **7**, 619–632 (2013).
25. Giles, H., Scherer, K. R. & Taylor, D. M. Speech markers in social interaction. in *Social markers in speech* 343 (Cambridge University Press, 1979).
26. Abitbol, J. L., Karsai, M., Magué, J.-P., Chevrot, J.-P. & Fleury, E. Socioeconomic dependencies of linguistic patterns in Twitter: a multivariate analysis. in *Proceedings of the 2018 World Wide Web Conference* 1125–1134 (International World Wide Web Conferences Steering Committee, 2018). https://doi.org/10.1145/3178876.3186011.
27. Hart, B. & Risley, T. R. *Meaningful Differences in the Everyday Experience of Young American Children*. (Brookes Publishing Company, Inc, 1995).
28. Gilkerson, J. et al. Mapping the early language environment using all-day recordings and automated analysis. *Am. J. Speech Lang. Pathol.* **26**, 248–265 (2017).
29. Broeck, J. V. D. Class differences in syntactic complexity in the Flemish town of Maaseik. *Lang. Soc.* **6**, 149–181 (1977).
30. Buck, B., Minor, K. S. & Lysaker, P. H. Differential lexical correlates of social cognition and metacognition in schizophrenia; a study of spontaneously-generated life narratives. *Compr. Psychiatry* **58**, 138–145 (2015).
31. Mota, N. B., Sigman, M., Cecchi, G., Copelli, M. & Ribeiro, S. The maturation of speech structure in psychosis is resistant to formal education. *Npj Schizophr.* **4**, 1–10 (2018).
32. Toth, S. L., Pickreign Stronach, E., Rogosch, F. A., Caplan, R. & Cicchetti, D. Illogical thinking and thought disorder in maltreated children. *J. Am. Acad. Child Adolesc. Psychiatry* **50**, 659–668 (2011).
33. Hough, S. D. & Kaczmarek, L. Language and reading outcomes in young children adopted from Eastern European Orphanages. *J. Early Interv.* **33**, 51–74 (2011).
34. Windsor, J., Moraru, A., Nelson, C. A., Fox, N. A. & Zeanah, C. H. Effect of foster care on language learning at 8 years: findings from the Bucharest early intervention project. *J. Child Lang.* **40**, 605–627 (2013).
35. Kornilov, S. A. et al. Language outcomes in adults with a history of institutionalization: behavioral and neurophysiological characterization. *Sci. Rep.* **9**, 4252 (2019).
36. Roy, P. & Chiat, S. Teasing apart disadvantage from disorder: the case of poor language. in *Current Issues in Developmental Disorders* (ed. Marshall, C. R.) 125–150 (Psychology Press, 2013).
37. Mackinley, M., Chan, J., Ke, H., Dempster, K. & Palaniyappan, L. Linguistic determinants of formal thought disorder in first episode psychosis. *Early Interv. Psychiatry* **15**, 344–351 (2021).
38. Tang, S. X. et al. Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. *Npj Schizophr.* **7**, 1–8 (2021).
39. Ayer, A. et al. Formal thought disorder in first-episode psychosis. *Compr. Psychiatry* **70**, 209–215 (2016).
40. Çokal, D. et al. The language profile of formal thought disorder. *Npj Schizophr.* **4**, 1–8 (2018).
41. Silva, A., Limongi, R., MacKinley, M. & Palaniyappan, L. Small words that matter: linguistic style and conceptual disorganization in untreated first-episode Schizophrenia. *Schizophr. Bull. Open* **2**, sgab010 (2021).
42. Buck, B. & Penn, D. L. Lexical characteristics of emotional narratives in schizophrenia: relationships with symptoms, functioning, and social cognition. *J. Nerv. Ment. Dis.* **203**, 702–708 (2015).
43. Zvelebil, K. 1927-2009. Personal pronouns in Tamil and Dravidian. *Indo-Iran. J.* **6**, 65 (1962).
44. Rajendran, S. Parsing in tamil: present state of art. *Lang. India* **6**, 8 (2006).
45. Mede, E. & Gürel, A. Acquisition of English articles in early bilingualism. *EUROSLA Yearb.* **10**, 193–219 (2010).
46. Chan, A. Y. W. How much do Cantonese ESL learners know about the English article system? *System* **56**, 66–77 (2016).
47. Ochsenbauer, A.-K. & Engemann, H. The impact of typological factors in monolingual and bilingual first language acquisition: caused motion expressions in English and French. *Lang. Interact. Acquis.* **2**, 101–128 (2011).
48. Hendriks, H., Hickmann, M. & Demagny, A.-C. How adult English learners of French express caused motion: a comparison with English and French natives. *Acquis. Interact. En Lang. Étrangère* 15–41 (2008) https://doi.org/10.4000/aile.3973.
49. Blodgett, S. L. & O'Connor, B. Racial disparity in natural language processing: a case study of social media African-American English. *ArXiv170700061 Cs* (2017).
50. Hollis, C. Child and adolescent (juvenile onset) schizophrenia. a case control study of premorbid developmental impairments. *Br. J. Psychiatry J. Ment. Sci.* **166**, 489–495 (1995).
51. Nicolson, R. et al. Premorbid speech and language impairments in childhood-onset schizophrenia: association with risk factors. *Am. J. Psychiatry* **157**, 794–800 (2000).
52. Haimo, S. F. & Holzman, P. S. Thought disorder in schizophrenics and normal controls: social class and race differences. *J. Consult. Clin. Psychol.* **47**, 963–967 (1979).
53. Parrott, B. & Lewine, R. Socioeconomic status of origin and the clinical expression of Schizophrenia. *Schizophr. Res.* **75**, 417–424 (2005).
54. Nogueira, A. S. et al. Influence of migration on the thought process of individuals at ultra-high risk for psychosis. *Braz. J. Psychiatry* (2020) https://doi.org/10.1590/1516-4446-2019-0685.
55. Berg, A. O. et al. The impact of immigration and visible minority status on psychosis symptom profile. *Soc. Psychiatry Psychiatr. Epidemiol.* **49**, 1747–1757 (2014).
56. de Sousa, P., Spray, A., Sellwood, W. & Bentall, R. P. 'No man is an island'. Testing the specific role of social isolation in formal thought disorder. *Psychiatry Res.* **230**, 304–313 (2015).
57. Sousa, P., de, Sellwood, W., Griffiths, M. & Bentall, R. P. Disorganisation, thought disorder and socio-cognitive functioning in schizophrenia spectrum disorders. *Br. J. Psychiatry* **214**, 103–112 (2019).
58. Goulding, S. M., Chien, V. H. & Compton, M. T. Prevalence and correlates of school drop-out prior to initial treatment of nonaffective psychosis: further evidence suggesting a need for supported education. *Schizophr. Res.* **116**, 228 (2010).

59. de Sousa, P., Sellwood, W., Eldridge, A. & Bentall, R. P. The role of social isolation and social cognition in thought disorder. *Psychiatry Res.* **269**, 56–63 (2018).
60. Holz, N. E., Tost, H. & Meyer-Lindenberg, A. Resilience and the brain: a key role for regulatory circuits linked to social stress and support. *Mol. Psychiatry* **25**, 379–396 (2020).
61. Dowd, J. B., Simanek, A. M. & Aiello, A. E. Socio-economic status, cortisol and allostatic load: a review of the literature. *Int. J. Epidemiol.* **38**, 1297–1309 (2009).
62. Goodman, E., McEwen, B. S., Huang, B., Dolan, L. M. & Adler, N. E. Social inequalities in biomarkers of cardiovascular risk in adolescence. *Psychosom. Med.* **67**, 9–15 (2005).
63. Pitts, M. J. & Gallois, C. Social markers in language and speech. in *Oxford Research Encyclopedia of Psychology* (2019).
64. Nilep, C. "Code Switching" in sociocultural linguistics. *Colo. Res. Linguist*. (2006) https://doi.org/10.25810/hnq4-jv62.
65. Yang, Y. & Eisenstein, J. Overcoming language variation in sentiment analysis with social attention. *Trans. Assoc. Comput. Linguist*. **5**, 295–307 (2017).
66. Hitczenko, K., Cowan, H., Mittal, V. & Goldrick, M. Automated coherence measures fail to index thought disorder in individuals at risk for psychosis. in *Proceedings of the seventh workshop on computational linguistics and clinical psychology: improving access* 129–150 (Association for Computational Linguistics, 2021).
67. Koenecke, A. et al. Racial disparities in automated speech recognition. *Proc. Natl Acad. Sci.* **117**, 7684–7689 (2020).
68. Zhang, H., Lu, A. X., Abdalla, M., McDermott, M. & Ghassemi, M. Hurtful words: quantifying biases in clinical contextual word embeddings. *ArXiv200311515 Cs Stat* (2020).
69. Leptak, C. et al. What evidence do we need for biomarker qualification?. *Sci Transl Med*. **9**, 417 (2017).
70. Goldsack, J. C. et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs). *Npj Digit. Med.* **3**, 1–15 (2020).
71. Panch, T., Mattie, H. & Atun, R. Artificial intelligence and algorithmic bias: implications for health systems. *J. Glob. Health* **9**, 010318 (2019).
72. Chen, I. Y., Joshi, S. & Ghassemi, M. Treating health disparities with artificial intelligence. *Nat. Med.* **26**, 16–17 (2020).
73. Mothilal, R. K., Sharma, A. & Tan, C. Explaining machine learning classifiers through diverse counterfactual explanations. in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* 607–617 (Association for Computing Machinery, 2020). https://doi.org/10.1145/3351095.3372850.
74. Panch, T. et al. "Yes, but will it work for my patients?" Driving clinically relevant research with benchmark datasets. *Npj Digit. Med.* **3**, 1–4 (2020).
75. Roth, W. D. The multiple dimensions of race. *Ethn. Racial Stud.* **39**, 1310–1338 (2016).

ACKNOWLEDGEMENTS

The author's research activities are partly supported by Canadian Institutes of Health Research [CIHR] Foundation Grant (grant number 375104); Bucke Family Fund, The Chrysalis Foundation and The Arcangelo Rea Family Foundation (London, Ontario). The author acknowledges salary support from the Tanna Schulich Chair of Neuroscience and Mental Health, Schulich School of Medicine and Dentistry.

AUTHOR CONTRIBUTIONS

LP is the sole author of this work and confirms meeting the criteria for contribution and accountability for the final version.

COMPETING INTERESTS

The author reports personal fees from Otsuka Canada, Janssen Canada, SPMM Course Limited, UK, Canadian Psychiatric Association; book royalties from Oxford University Press; investigator-initiated educational grants from Janssen Canada, Sunovion and Otsuka Canada in the last 3 years, all outside the submitted work. The author is also the convenor of the Diverse International Scientific Consortium on Thought Language and Communication in Psychosis (DISCOURSE in Psychosis: https://discourseinpsychosis.org/).

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to L.P.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.