Schizophrenia Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

Schizophrenia Research

journal homepage: www.elsevier.com/locate/schres

Reflections on the nature of measurement in language-based automated assessments of patients' mental state and cognitive function

Peter W. Foltz [a,*], Chelsea Chandler [a,b], Catherine Diaz-Asper [c], Alex S. Cohen [d,e], Zachary Rodriguez [d,e], Terje B. Holmlund [f], Brita Elvevåg [f,g,**]

[a] Institute of Cognitive Science, University of Colorado Boulder, United States of America
[b] Department of Computer Science, University of Colorado Boulder, United States of America
[c] Department of Psychology, Marymount University, United States of America
[d] Department of Psychology, Louisiana State University, United States of America
[e] Center for Computation and Technology, Louisiana State University, United States of America
[f] Department of Clinical Medicine, University of Tromsø - the Arctic University of Norway, Tromsø, Norway
[g] Norwegian Centre for eHealth Research, University Hospital of North Norway, Tromsø, Norway

ARTICLE INFO

Keywords:
Natural language processing
Speech technologies
Artificial intelligence

ABSTRACT

Modern advances in computational language processing methods have enabled new approaches to the measurement of mental processes. However, the field has primarily focused on model accuracy in predicting performance on a task or a diagnostic category. Instead the field should be more focused on determining which computational analyses align best with the targeted neurocognitive/psychological functions that we want to assess. In this paper we reflect on two decades of experience with the application of language-based assessment to patients' mental state and cognitive function by addressing the questions of *what* we are measuring, *how* it should be measured and *why* we are measuring the phenomena. We address the questions by advocating for a principled framework for aligning computational models to the constructs being assessed and the tasks being used, as well as defining how those constructs relate to patient clinical states. We further examine the assumptions that go into the computational models and the effects that model design decisions may have on the accuracy, bias and generalizability of models for assessing clinical states. Finally, we describe how this principled approach can further the goal of transitioning language-based computational assessments to part of clinical practice while gaining the trust of critical stakeholders.

1. Introduction

Our verbal expressions provide a unique lens to our inner thought processes and thereby an indirect window into the brain and potential pathologies. 'Distortions' in language are medical *signs* that are measurable, but at present no measurement is universally accepted. *"An analogy to the equally noninvasive thermometer is that language provides an index into processes inside the body. An abnormal temperature from a cold does not indicate a disease of temperature regulation, but rather provides an indirect pathway toward measuring the internal processes contributing to the observed deviation. In the case of language measures need to be established and calibrated"* (p. 510; Elvevåg et al., 2017).

Twenty-five years ago we embraced the opportunity afforded by the evolving advances in computing power and algorithms to analyze speech from patients with schizophrenia using the latest natural language processing and machine learning methods in order to create new language measures. We hypothesized that these approaches could result in measures that were potentially more useful than those traditionally available. We designed a fairly simple study to collect free speech samples from a variety of tasks that would elicit speech in a modest sample size (26 patients with schizophrenia as compared to 25 age-matched healthy controls) (Elvevåg et al., 2007). The study adopted the popular methods of the time of collecting data once only (i.e., a cross-sectional snapshot in time) using an in-person controlled assessment method (i.e., a skilled interviewer physically present in the same controlled lab setting as the participant) and compared the novel metrics

* Correspondence to: P.W. Foltz, Institute of Cognitive Science, University of Colorado Boulder, United States of America.
** Correspondence to: B. Elvevåg, Department of Clinical Medicine, University of Tromsø - the Arctic University of Norway, Tromsø, Norway.
E-mail addresses: peter.foltz@colorado.edu (P.W. Foltz), brita.elvevag@uit.no (B. Elvevåg).

https://doi.org/10.1016/j.schres.2022.07.011
Received 31 March 2022; Received in revised form 12 July 2022; Accepted 13 July 2022
0920-9964/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

- derived using natural language processing (NLP) (e.g., Jurafsky & Martin, 2008) and early word embedding methods (e.g., Foltz, 1996; Landauer et al., 1998) - to clinical ratings of these patients that were collected within a few days of the experimental study. Our overarching goal was to use these methods to improve current assessment methods in psychiatry (in terms of objectivity and sensitivity) with an approach that was very similar to how student writing was computationally assessed in the education domain (Foltz et al., 1999). We applied this analysis approach to clinical interviews (i) to provide a second opinion about illness severity (notably severity of thought disorder) and (ii) to understand the underlying neurocognitive mechanisms (specifically memory) of disordered thinking in clinical conditions that affect cortical function.

We found that we could represent participant's verbal responses as vectors in a high-dimensional semantic space and quantify linguistic aspects of their responses as changes in these vectors over their response. This semantic quantification was enhanced with the use of machine learning to combine it with other NLP-based measures that characterized the complexity and sophistication of the syntactic and lexical features of patients' language. This approach provided the basis to derive a range of measures including participants' coherence, tangentiality and amount of relevant content in their responses. Across a number of assessment types (e.g., word association, verbal fluency, storytelling) we found that the measures agreed strongly with clinician assessments, could predict diagnostic categories and provided a framework for analyzing discourse to understand the nature of disordered language production. While the approach was initially met with a large degree of skepticism by reviewers and editors, the eventual publication was the very first to apply modern high dimensional language-based methods to psychiatric settings.

Our initial work of improving assessment and understanding better the nature of the clinical presentation of patients by differentiating those with schizophrenia from healthy controls was then extended by applying similar methods to discriminate schizophrenia probands, first-degree relatives and unrelated healthy controls (Elvevåg et al., 2010), to differentiate those at high risk of psychosis from unrelated putatively healthy participants (Rosenstein et al., 2015a), as well as try to understand speech coherence better by linking language to underlying neurobiology in a candidate gene study (Nicodemus et al., 2014) and by brain imaging (fMRI) (Tagamets et al., 2014). To align with our goal of leveraging NLP methods to improve assessment, we built on this general methodology to enable remote monitoring of psychiatric outpatients and moved - before the covid19 pandemic - this type of assessment out of controlled settings such that tests could be self-administered remotely (Holmlund et al., 2019a,b), thus leveraging technology to complement, change and potentially disrupt how language-based assessment has historically been conducted in psychiatry.

Concurrently, we sought to leverage these methods to better understand the nature of the memory problems associated with schizophrenia by focusing on story recall in a verbal recall subtest of a very widely used neuropsychological test. Verbal memory deficits are a hallmark feature in schizophrenia (Heinrichs and Zakzanis, 1998) and so it stands to reason that modeling aspects of language features used in recall could measure these verbal processes. We found that automated analyses of the recalls accurately mimicked human scoring, and as in our previous work (e.g., Elvevåg et al., 2007, 2010), the semantic features were most predictive and able to differentiate patients, their unaffected siblings and healthy controls (Rosenstein et al., 2014). Thus, the same NLP technology was used to automate human approaches to assessing the nature of memory deficits as evidenced via verbal recall tasks. Since stories are fundamental to the human experience and provide an effective way to organize information, evaluating how stories are recalled provides critical information about the health of our brain, and we found that speech technology can successfully automate such evaluations (Chandler et al., 2019; Holmlund et al., 2020a).

Over the past two decades, this area has continued to grow, both by our research group as well as by many other researchers (e.g., Bedi et al., 2015; Fraser et al., 2015; Morgan et al., 2021; Rezaii et al., 2019; Voppel et al., 2021; for reviews and workshops, see Corcoran and Cecchi, 2020; Goharian et al., 2021; Low et al., 2020). The field has examined both structured language, such as verbal recalls, as well as more unstructured language such as those generated in social media posts and personal diaries, and has been evaluated on data from a wide range of clinical areas including schizophrenia, cognitive decline, suicidality and depression. The growth in the field is enabled by continuing advances in AI-based language methods, increased computing power, the ubiquity of mobile devices for data collection and greater recognition of how such approaches can isolate the underlying mechanisms of serious mental illnesses (e.g., Cohen et al., 2017; Elvevåg et al., 2016). In keeping with the integrative and dimensional transdiagnostic approach of the NIMH Research Domain Criteria to compare across disorders, at some level, this work of creating new language measures for assessment that specifically leverage NLP methods can be conceptualized as one of the critical components of the foundational framework that may help to eventually (re-)bridge the two seemingly disparate medical specialties of psychiatry and behavioral neurology. As much of neurocognition and mental state is assessed through speech and language, leveraging NLP methods to operationalize language constructs may help to link the brain, mental states and verbal behaviors (Martin, 2002; Yudofsky and Hales, 2002). The approach further imparts clinical translation value with the means for providing rapid, accurate evaluations, which enables novel ways of scaling assessment and new models of monitoring patient states.

Overall, the methods we devised in our original study were sound and pointing in the right direction. However, through our subsequent work, we have learned that there are many more considerations that must be made in designing effective approaches to measurement, both for effective assessment and for clinical transition. We are therefore extremely grateful to have the opportunity to share our reflections on the state of the use of these computational language approaches for measurement in psychiatry.

For our reflections, we loosely adopt notions used in Evidence-centered assessment design (ECD) (Mislevy et al., 2003). ECD is a framework for developing assessments that considers the evidentiary arguments that support the validity claims. In essence, it seeks to link the kinds of tasks that elicit language behaviors (e.g., assessments) to the evidence derived by the measures (e.g., NLP-based analyses and rubrics) and to the target neuropsychological constructs (e.g., clinical states). The approach helps to assure the validity of the assessments as well as provide transparency and accountability for the evidentiary reasoning that is derived from the assessment approach. We use the structure from ECD to reflect on the nature of *what* we are measuring, *how* we are measuring it and *why* we are measuring it. While we discuss this framework in the context of specific examples of our research, our goal is to provide an approach that is data agnostic, in that it should not matter what symptoms, scales and computational methods are used, but provides a framework for thinking about the *what*, *how* and *why* of computational language assessment.

2. What are we measuring?

> *"In general, we inherit the questions of our intellectual predecessors, who knew even less than we do, and thus risk seeking explanations for concepts that were not defined in a manner that best captures the real processes of interest. Consequently, we are motivated to periodically re-examine the questions we aim to answer and look outside whatever field we have defined ourselves into"*
>
> (Cisek, 2019; p.2265)

2.1. Clinical constructs

The first issue we address is *what is the collection of patient attributes that should be assessed*? Put differently, is it the clinical states, the neurocognitive performance profile, the daily fluctuations in some assay or something else entirely? This can be conceptualized as the *clinical states model*. Two centuries ago, measurements of the (*static*) bumps on the skull - which were quantified with the methods of phrenology - were central attributes for the understanding of risk and propensity to deviant behavior. A century later, mental states were conceptualized using the new phenomenological framework of Bleuler (1911, 1950) that could embrace their *fluctuating* nature. A few decades ago, *putatively static* neuropsychological constructs were of core interest and were employed in the quest to understand genetic risk of mental disorder by defining putatively stable intermediate phenotypes (intermediate in that they bridge the gap between effects that genes have at the cellular level to the emergent psychosis), which at the level of behavior could be a heritable trait in performance in a neurocognitive domain. Indeed, even after the introduction of brain scanning and genetics in psychiatric research, most assessment remains through the medium of language at some level, and a large range of neurocognitive function and thus deficits are measured through language, as well as of course clinical state and the charting of the verbal medical signs (e.g., of language – for early modeling of this discourse see Hoffman et al., 1986; Hoffman, 1987). Today - courtesy of technological advances in mobile devices and wearable sensors - there is a growing awareness that assessment can go beyond single cross-sectional snap-shots in time to look at multiple measurements in order to model the *dynamics* of the clinical state (Cohen et al., 2019; Holmlund et al., 2020b; Ranjan et al., 2022). These assays include both neurocognitive and clinical measures, and no longer assume cognition is static and clinical states are fluctuating (i.e., the trait-state distinction is flawed; Cohen et al., 2019; Cohen et al., 2021; Cowan et al., 2019; Le et al., 2021).

However, do we really know what collection of attributes of the patient should be assessed? Should the attributes of interest be the phenomenological expressions of clinical state or the underlying neurocognitive processes, or can it be both? At present it is a mix of measures that are static, fluctuating, single attributes or a selection of the aforementioned, and there is little agreement on how they should all be weighted (e.g., the same attribute could have a test that focuses on the stability versus fluctuating nature of the construct as it is a property of how the test is developed). What we do know is that with more frequent measurements, we start to see magnitudes more instability, and at present lack the norms for interpreting the sheer volume of data, different channels and data types (Chandler et al., 2020b). Assessment batteries have generally been designed to elicit these deficits, and automating this assessment process allows and forces precision in our definition (which thus changes the definition) of what aspects of language we are specifically measuring. Further, modeling all of this data necessitates a new framework, namely dynamical psychometrics (Cohen et al., 2020a,b; Cohen et al., 2021).

2.2. Evidence models

The second issue we address is *what should behaviors reveal at the different levels of the targeted states?* To address this we examine the nature of certain behavioral outputs to determine the nature of the disorder. This of course is affected by the nature of the assessment which may be conceptualized as the *evidence model*.

Consider the case of the semantic verbal fluency task where participants are asked to name as many words that fall in a particular semantic category (e.g., animals) within a minute. This type of task is probably the most widely used verbal assay of executive function/semantic memory within the field of neuropsychology (Lezak et al., 2012; Strauss et al., 2006) and is widely used in psychosis research. Nevertheless, the most common hand-scoring approach to assessing semantic fluency is to count how many exemplars are generated. This misses rich information about the semantic and temporal structure that may reflect a responder's mental state (as examined in Elvevåg et al., 2007). A recent study of ours applied both speech processing and NLP approaches to the verbal output from this traditional category fluency task and showcased how the verbal output process can be informed in detail both semantically and temporally (Holmlund et al., 2019a,b). Thus, these NLP techniques allow us to move beyond simple counting of single utterances and groups of related utterances (e.g., categories; Troyer, 2000). A semantic word embedding space can be computed in a specific domain (e.g., animals) in order to extract fine-grained relatedness measures of examples as well as determine related clusters and how quickly a participant moves from cluster to cluster (e.g., Rosenstein et al., 2015b). This approach can be helpful clinically to better inform symptom definitions (e.g., coherence). It further has the added benefit that exemplars not be grouped and labeled by hand, thus minimizing human biases (but note that human biases can still occur in NLP-based models because of biases inherent to the actual text corpora that are used to create such models - see e.g., Hitczenko et al., 2022).

Beyond showing that these features improve predictive performance of machine learning models (e.g., distinguishing patients from controls), it is critical to link the creation and use of these features to the constructs being assessed. Thus, these semantic word embedding and temporal features force a discussion on their evidentiary role. For example, the development of features such as counts of words recalled may reveal slowing in executive function (e.g., Ghanavati et al., 2019), measures of temporal switching between clusters may be related to working memory capacity (Oh et al., 2019) and similarity between words may indicate cortical structures (although note that these may not reflect structural differences in semantic memory - Voorspoels et al., 2014).

The evidentiary process provides a basis for inferencing about how the computational features are implicated in verbal behavior. However, it is equally important that the features are *validated* across data sets of varying clinical conditions in order to establish their effectiveness in assessing the constructs of interest in a way that predicted results from the model are *interpretable* and *generalizable*. For such a process, we appeal to the field of language testing which has worked on developing arguments to validate the alignment of language constructs to the resulting scores from assessments. Derived from work from Kane (1992), the field defines *validation* through a systematic process of developing interpretive arguments with which to specify the *constructs*, intended *decisions* and *consequences* of the resulting scores from model predictions (e.g., Bachman, 2005; Bachman and Palmer 2010; Chapelle et al., 2010). Using such a validity argument approach provides a framework that shows how the inferences provide support for the interpretive argument (e.g., Chapelle, 2012). This approach has been applied to health-based assessments (e.g., Hawkins et al., 2021), and can easily be adapted for language features for assessing mental health. Table 1 below shows an example of how it can be applied for NLP features for the semantic verbal fluency task incorporating the computational assumptions made by the use of features. Using this inferential process helps clarify *how the constructs are instantiated* in the specific assessment, how the NLP-based model is *evaluated*, *how it should generalize* across tasks and/or related data, *how well the construct can be extrapolated to other criteria* and finally, *how the assumptions and predictions of the machine learning (ML) model support a clinician's decision-making*. Clinical assessment is ultimately about gathering evidence to support claims we wish to make about patients. However, at present, most research on computational models of mental health seldom address the issues inherent in the validation component of the inferential process and instead focus on reporting just at the *evaluation* level of how well a specific model agrees with a diagnostic category.

Table 1
Sample inferential process for a semantic verbal fluency task.

Inference	Basis for inference	Example assumptions underlying inference	Computational assumptions
Domain description	Observed prediction tells a story about cognitive/psychological state in situations in the target domain	Semantic verbal fluency assessment indicates breakdowns of semantic structures related to neurocognitive states	Apply NLP features that measure semantic distances and clusters
Evaluation	Observed prediction reflects targeted psychological state	Statistical characteristics of items, measures and assessment forms are appropriate for clinician decisions	ML model predicts clinicians' rating of fluency quality or overall diagnosis
Generalization	Observed prediction provides similar predictions as parallel tasks or assessments	The test includes a sufficient number of tasks to provide stable estimates of test taker performance	ML model predictions are accurate across parallel task types (e.g., other semantic fluency task versions)
Explanation	Expected scores are attributed to a target psychological construct	The internal structure of the test score is consistent with a theoretical view of the neuropsychological construct	ML model predictions are attributed to relevant constructs
Extrapolation	The constructs being assessed by the test account for changes in cognitive/psychological state	Assessment performance is related to external criteria	ML predictions have associations with relevant external criteria (e.g., predict diagnostic category)
Utilization	Performance estimates on the assessment are useful to clinicians for making diagnostic decisions	Clinicians can easily interpret scores	ML predictions support high-stakes decisions

Table 2
Examples of constructs that can be extracted from a story recall with associated computational measures.

Story recall construct	Computational feature	Example References
Overall amount of relevant information recalled. Number of semantic concepts recalled.	word2vec word movers distance of recall to original story	Chandler et al. (2019)
Narrative structure	Distribution of key entity frequencies with a Markov Chain model for transitions	Chaspari et al., (2013), Prud'hommeaux and Roark (2015)
Use of proper grammatical/syntactic structure	Parse tree depth and alignment to original story	Roark et al. (2011)
Memory decay over recalls	word2vec cosine distance of immediate recall to recall delayed by 1 day	Chandler et al. (2019)
Story Coherence	LSA cosine of 5 word window to next window	Elvevåg et al. (2007), Iter et al. (2018)
Tangentiality	Slope of cosines of word windows that are 1 to n word distant from the original window	Elvevåg et al. (2007), Morgan et al. (2021)

2.3. Construct driven NLP features aligned to assessment tasks

As a further example, we examine tasks and features used in assessing verbal recollections. Indeed, at the core of the patient-clinician interaction is the *anamnesis*, a medical term derived from the Greek words 'open' and 'memory', and thus in essence meaning 'reminiscence'. This process of taking the medical history of patients involves the clinician asking questions either directly to the patient to probe from memory or to those who know the patient well enough to recall information considered useful (by the clinician) in terms of diagnosis and treatment of that patient. The questions are intended to generate a recollection from memory of information about the symptoms. Additionally, in the case of psychiatry these questions are designed to prompt a fairly detailed personal life history. Indeed, the process of this personal storytelling is paramount to a correct diagnosis and the story generated affects subsequent care management plans.

At first glance, such detailed and phenomenological medical case histories might seem at odds with reductionist approaches to research, yet it is arguably these very stories that may both provide the bridge between phenomenology and reductionism as well as contain the critical clues regarding levels of wellness or function. In part this is because although traditionally the stories are viewed as a method to elicit information regarding *symptoms*, if formalized using reductionistic methods may also be viewed as information about medical *signs*.

The questions that the clinician asks requires that the patient understands the question, a process that involves numerous components (e.g., extracting the words, decoding the propositions, contextualizing) and when the patient responds they need to decide on a response (presumably after searching through memory), construct phrases, select the lexical items, build the clause structure and articulate the response (see e.g., Levelt, 1989). Reductionistic behavioral approaches have traditionally attempted to measure all these underlying components by creating test batteries that purport to test these constructs (e.g., attention, verbal memory, working memory) through individual specialized assessments. However, it is also possible to obtain these very constructs from the storytelling process. Critical to measurement though, is to align these constructs to computational features. Table 2 shows such an alignment for several constructs. By explicitly defining how each feature aligns to different constructs, we can then simultaneously extract evidence for each construct.

Taken together, we see that there needs to be a strong interrelationship and inter-dependency of the psychological construct, the task used to elicit the speech and the computational analysis which provides the putative evidence of the inferred clinical states that provides actionable information for the clinician. The patient's brain mediates language/speech which is elicited by tasks and analyzed by computational methods which provide the evidence. In Fig. 1, we illustrate this relationship between these factors with an example of a story recall task that requires the participant to retell the original story and is analyzed by computing the average cosine of moving windows of word2vec cosine distances to consecutive windows, and the level of deviation compared to that of the original story, and is used to assay coherence which is inferred as reflecting the current cognitive and mental state of the patient.

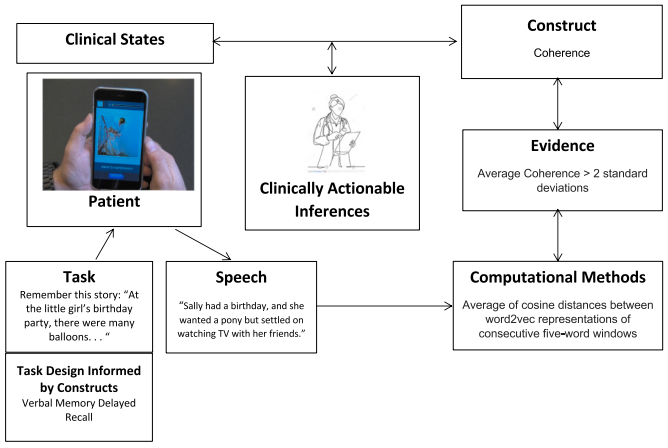

Fig. 1. Interrelationship of task, computational methods and evidence for psychological constructs providing actionable evidence for clinicians. Double arrows indicate how the methods, evidence, constructs and clinical states are all linked in both directions.

2.4. Dynamics in measurement

Most neuropsychological assessments examine patient states at a single point in time. However, there is a pressing need for research frameworks that embrace language technologies to also include temporal dynamics and the quality of individuals' integration of information in order to produce verbal behavior (e.g., Cohen et al., 2021). Furthermore, in any application of behavioral or neuropsychological test, predicting risk and the time scale of *when* this clinical event or significant change will happen is critical. To date, the actual temporal nature of this has not been formally addressed in the existing literature (Holmlund et al., 2021). Thus, seemingly impressive NLP findings in research in high risk youth identifies a signal that has predictive value several years later (e.g., Bedi et al., 2015; Corcoran et al., 2018), and likewise in mild cognitive impairment (MCI) finds a subtle signal in language usage that predicts an elevated risk of developing dementia several years later (e.g., Eyigoz et al., 2020). Yet are these signals the same across illnesses *and* also time? Are the signals that are evident in minutes before an individual plunges into depression and feels suicidal similar or different to those evident in other conditions in which the time course is radically different? The answer to these questions will help us understand what exactly these language signals are, how stable they are and what factors influence their change. Put differently, what does the temporal dimension tell us about the features we are observing (e.g., a stable measure in a behavioral trait or a fluctuating state issue), and can we use the NLP output to look at the features to establish what it is that this underlying signal is coming from? From a practical perspective, this issue of differences in prediction time frames between various studies stems from the targets in the machine learning system themselves rather than if a specific feature X can predict illness a year out versus whether a specific feature Y can predict illness within an hour. This may be the case in the real-world settings, but in machine learning models, algorithms can only learn what they are given. If the gold standard labels are such that they were given from a certain time frame, then the model will only be able to learn in the time frame. Therefore, it is critical to select training data that is representative of the time frame and scale that will be used for analysis.

But how much detail is necessary for a dynamic model? Often it is argued that increasing the level of detail at the psychometric level will be beneficial (e.g., to uncover the effect of a functional polymorphism on functional and neurocognitive phenotypes - Elvevåg and Weinberger, 2009). Historically, the studied measures have been opportunistic based upon available data and simply computed the average differences between groups of people, and the relative slower or poorer performance, which are obviously an oversimplification of behavior patterns that ignore individual variability (and intra-individual variability). However, although it is logical to assume that capturing fluctuations - that may reflect transient or enduring changes - in behavior (e.g., variations in performance day-to-day or trial-to-trial) will be useful clinically (especially in terms of our understanding of pharmacodynamics influencing how a patient responds to a specific medication), it remains an empirical question and the answer is likely context specific and task specific. Indeed, the incorporation of real-time data collection in models to chart the fine-grained temporal nature of cognitive and mental states and their interaction will necessitate dynamical psychometrics and a dynamical cognitive neuroscience approach. Again, although it is intuitive that the future success in terms of contributions to neuropsychiatry will require combining realistic time constants at all levels of cognitive neuroscience (molecules, neural systems and cognition) this remains to be established empirically. Addressing such questions can advance both theoretical notions of states of mental illness as well as empirical measures that allow us to better quantify how we define and measure an individual's mental state and the importance of detecting the changes in state.

We see dynamics as a challenge and growth area for the field of NLP in mental health, requiring more systematic data collection and development and testing of new methods that combine the signals derived from the NLP measures with methods that measure the important changes and account for the contexts needed to be able to interpret the changes in individuals' states. From a data perspective, longitudinal methods require many data points, necessitating more regular sampling of language from individuals. While structured tasks can be administered regularly, language from unstructured data (e.g., tweets, emails) may be more naturalistically captured. From a methodological perspective, the field has often focused on comparing an individual's state to population norms. Longitudinal measurement provides the opportunity (assuming enough data), to treat individuals as their own baseline, and it is the change from that state that is important to measure. To this end, researchers employing NLP should be looking to other areas where longitudinal modeling has been applied. For example, Experience Sampling Methods/Ecological Momentary Assessment (ESM/EMA) (e.g., Trull and Ebner-Priemer, 2009) provide a framework for realtime data collection of assessments with a focus on capturing and combining data from an individual's state at multiple time points. It has thus far only been applied to more typical mental health assessments, but seems primed for the incorporation of NLP-based measures and novel tasks. Similarly, there has been growing work on applying neural network models (e.g., convolutional neural networks and long short-term memory models) on a range of longitudinal medical data (e.g., EHR records to predict cardiovascular events - Zhao et al., 2019; detection of change through radiological exams - Santeramo et al., 2018).

3. How are we measuring?

> *"The nature of the construct being assessed should guide the selection or construction of relevant tasks, as well as the rational development of construct-based scoring criteria and rubrics."*
>
> Messick, 1994, p. 20

Next we examine how measurement can be done through tasks and computational methods while considering fidelity to the constructs. Applying novel computational methods to patient speech affords us the opportunity to re-think the kinds of tasks that are used to elicit verbal response and evaluate their effectiveness. Concurrently, we also need to examine the assumptions underlying the computational methods being used for the tasks to ensure that they provide accurate, unbiased assessments.

3.1. What tasks are best?

We first consider what tasks are best for eliciting the target behaviors. Much of the research in computational psychiatry to date has largely focused on applying artificial intelligence techniques to standard neuropsychological tasks (e.g., Wechsler Memory story recall subtest, semantic verbal fluency, cookie theft picture test). This has the benefit of years of research supporting that these specific tasks elicit the nuances in patient responses that allow clinicians to diagnose various disorders. Furthermore, clinicians are already familiar with their use, allowing for data collection in standard clinical settings. However, these standard tasks were never designed for the application of artificial intelligence methods. Their associated scoring rubrics are often designed for test administrators to simply count the units in patient responses and do not necessarily elicit measures of more fine-grained symptoms that may arise from patient speech. As such, there is great need to design new tasks that can be optimized for the collection of a broad range of constructs and modalities which artificial intelligence techniques are well suited to analyze. The evolution of NLP and machine learning techniques beyond simple counting warrants a complementary evolution in language-based neuropsychological tasks.

As computers are able to process and understand increasingly large amounts of data, NLP techniques are ideal to be the 'eyes that peer through the window' into the mechanisms underlying a person's neurocognitive processing. A central goal with NLP has always been to train a computer to "understand" language in a manner similar to humans. As such, it has evolved from simple rule-based understandings, where an algorithm is based solely on direct matching of input text to "if, then" statements, not adding any additional insights beyond what has been supplied by the programmer, to deep neural networks that have been trained on massive amounts of human language such that it can generate its own insights based on an extrapolation of knowledge from many potentially disparate sources. While we cannot claim that NLP is *interpreting* language in the same ways as humans, the predictions produced on assessing open-ended language in many different tasks and domains can sufficiently match those of human ratings to be highly useful for clinical assessment (e.g., Corcoran and Cecchi, 2020).

While prior neuropsychological assessments have been designed to strongly control the range of potential language responses (e.g., repeat a story), NLP provides the ability to introduce more open-ended naturalistic tasks and/or collect language in the uncontrolled wild. Thus, tasks that ask for a narrative of a person's day, a process ('tell me how you do your laundry'), picture descriptions or movie narrations provide some level of structure and concreteness while also providing personal relevance to engage a participant into providing a sufficiently large language sample. Concurrently, such tasks can assess multiple constructs simultaneously, such as semantics, grammatical structure, word frequency, sentence and phrase complexity and coherence, thereby providing a more nuanced and fine-grained analysis of neuropsychological functioning.

At the other end of the spectrum of openness, language obtained from unstructured communication and social media (e.g., Twitter or Facebook feeds, phone recordings, diaries) can often provide very large samples of data (e.g., Clarke et al., 2020; Coppersmith et al., 2014; De Choudhury et al., 2016; Guntuku et al., 2017) which can generate more general models of the relationship of language features to classes of mental illness. These can provide research insights into how different language features align to constructs as well as help in building systems to detect important changes in broad samples. However, these sources have very little control of the kinds and structure of language that is generated and so while good for characterizing overall population differences, they may be less accurate for characterization at the individual level. Nevertheless, in all cases of analyzing language data, it is critical to observe the purpose or task that drove the generation of the data and how that drives the language cues tied to the features and constructs used for analyses. In designing such tasks, it is critical to ensure that they elicit language that illustrates the constructs of interests. This requires testing of tasks and assessment models to assure that constructs are present in a way that is measurable and clinically useful. As such, this cannot be done solely by a machine learning/NLP expert, or a clinician, or a neuroscientist, or an assessment specialist, but benefits most from an iterative development process where the expertise from each is combined to provide an effective task and method for assessing the output.

3.2. Task creation with NLP in mind

Automating analyses further opens the possibility of developing many more forms of a task. For example, formal assessment of the verbal recall process is a core component of neuropsychological test batteries, the assays derived are of some of the most promising intermediate phenotypes in psychiatry and NLP assays obtained on speech including story recall are considered possible candidates for biomarker development (for review see, Corcoran and Cecchi, 2020). In the case of verbal memory recall, the commercial availability and hence dominance of the Wechsler Memory Scale has ensured that globally the vast majority of assessments of story recall (from the Logical Memory subtest) are remarkably enough based upon the recall of only two stories (of 65 and 86 word lengths; Wechsler, 2009). Yet it is easy to design new stories, which arguably may provide improved contextual relevance, yet it still needs to be established whether the novelty of the new stories has any real advantage over existing ones in terms of revealing an intermediate phenotype that provides a useful clinical target. Previously, we designed 24 variants of stories (with a range of 69 to 82 words in length) for remote administration and showed that it was possible to use NLP to successfully rate the story recall in a manner similar to trained experts (Chandler et al., 2019; Holmlund et al., 2020a).

Drawing from NLP analyses of these numerous story versions, Chandler et al. (2021a,b) further showed how NLP can inform the future design of the actual stories used to elicit behavior by ensuring that there is accurate machine learning prediction of the expert human ratings, and that it is possible to generate reliable predictions over time and over alternate forms of the same test in healthy individuals. Indeed, it stands to reason that the collection of responses to various verbal memory prompts can inform us in the choice of the optimal story that is to be remembered. For instance, when comparing hypothetical story variation A with story variation B, we can analyze the recall responses in a large dataset and find that performance can be better scored with machine learning models on story A than story B (because of closer alignment of the NLP features to the goals of the recall task) and that individuals consistently score in the same percentile of their population group when completing story A, but with story B participants are placed in alternate percentiles. Thus, we can use this approach to analysis to conclude that story A is well suited to providing highly specific and reliable assays of verbal memory. Undoubtedly this approach can be applied to the generation of alternate forms for other neuropsychological tasks also.

3.3. Assumptions underlying semantic measurement models

While some natural language processing techniques are viewed as a black box, an understanding of the processes with which they are governed is necessary in determining which methods are best suited for what types of neuropsychological tasks and the associated patient responses, why certain algorithms differ in terms of output from one another and critically how to appropriately apply the methods in varied scenarios. Each type of NLP method carries their own set of assumptions that must be considered in the design. As an example, an overview of the evolution of semantic models and their application to various tasks used in Elvevåg et al. (2007) is given.

(1) The earliest vector space models of semantic understanding were based on distributional properties of language - namely, latent

semantic analysis (LSA), which is a process that applies a matrix factorization to a large matrix of word co-occurrences. As the resulting word vectors from this approach are specifically generated from a co-occurrence matrix factorization, the LSA process entails a very specific understanding of single words and the contexts in which they tend to occur. However, the standard LSA model was built on a relatively small and restrained corpus of language (TASA; Landauer et al., 1998), resulting in many out of vocabulary instances and a constrained cultural and semantic space.

(2) Word embeddings saw multiple updates before the introduction of word2vec in 2013, however it was the work of Mikolov et al. (2013) where neural networks became the popularized approach to the creation of a robust semantic space. These embeddings were created by training a model to predict a word given the context in which it is used. Subsequently, Pennington et al. (2014) created GloVe embeddings in a fashion more similar to that of LSA with word co-occurrences. Both approaches result in single non-contextualized word vectors for each unique token (for instance, the vector for the word "check" would be the same in "I went to the doctor for a check up" versus "I cashed the check at the bank").

(3) To address this issue of non-contextualized word embeddings, language models became the mainstream way to represent language in a contextualized manner. In this class of approaches, it is possible to retrieve a distinct word embedding for an individual word (and additionally for entire sentences or paragraphs) when used in differing contexts. With the introduction of ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), USE (Cer et al., 2018), not only did word embeddings become contextualized, but it became possible for entire utterances to be encoded in a unique manner to the context.

With these newer and varied approaches, to what degree do they change performance in clinical tasks? As part of a novel investigation of word embeddings, we next compare these different word embedding approaches on the varied speech elicitation tasks from Elvevåg et al. (2007) to ascertain which approaches may be more appropriate for differing types of language. We chose to use the original data in order to determine the degree to which these semantic models have improved over the years, although there are other studies which have examined semantic models with newer data sets (e.g., Chandler et al., 2019; Iter et al., 2018; Rezaii et al., 2019). Specifically, we compared (1) the LSA word embedding model trained on the TASA corpus (http://lsa.colorado.edu/), (2) the word2vec model trained on the Google News dataset (https://code.google.com/archive/p/word2vec/), (3) the GloVe model trained on the combined Wikipedia 2014 + Gigaword 5th Edition corpora (https://www.kaggle.com/datasets/rtatman/glove-global-vectors-for-word-representation), (4) the USE model accessed via Tensorflow Hub (https://tfhub.dev/google/universal-sentence-encoder/4), and (5) the BERT-Base model accessed via Tensorflow Hub (https://tfhub.dev/tensorflow/bert_en_uncased_preprocess/3).

We strove for a more nuanced understanding of which embedding techniques are more fitting for different types of language and urge researchers to *jointly* consider the assumptions that went into the design of the natural language processing tools, as well as those that went into the design of the elicitation tasks, and how these might interact with one another. The assumption is that certain types of embedding approaches will be better suited to measure certain symptom classes, diagnostic categories, and (certainly) speech elicitation tasks. While the data set of Elvevåg et al. (2007) is limited in terms of participant size and its unidimensionality (i.e., simply composed of the transcribed text of participant responses to tasks), nevertheless it is an appropriate dataset for comparing various semantic space techniques as it is composed of a wide range of both classic and novel tasks. A description of each task is given along with the features we seek to measure and how appropriate each semantic NLP approach listed above is for the understanding of the particular nuances in language for the distinction between healthy and disordered speech.

The word association task was the first investigated in this comparison. The participants were presented with a set of individual words and were asked to recite the first word that came to mind. Some examples of the source words are: *bread, friends, king, sports, table*. The methodology applied to this task for differentiating patient class was to compute a direct cosine distance between the source word and the patient response using LSA, word2vec, GloVe, USE and BERT embeddings. F statistics were computed between the average cosine distances of the patient and control class as a means to ascertain which method created more significant separations between the two classes. In the task of individual word comparisons, LSA significantly outperformed other approaches (LSA $F = 7.57$, $p = 0.006$, word2vec: $F = 3.49$, $p = 0.06$, GloVe $F = 4.38$, $p = 0.04$, USE: $F = 4.66$, $p = 0.03$, BERT: $F = 0.14$, $p = 0.71$). Certainly, contextual information from surrounding language is not needed to understand direct relationships between words. As LSA does not specifically encode any information about the order of words in language and simply seeks to understand - at a word level - how words tend to appear in similar semantic contexts, it may be viewed as analogous to the task that is being measured with word associations. The constrained semantic space of LSA allows the resulting word embeddings to embed more concrete meanings for each word than the more modern approaches and the lack of proper language in the task does not necessitate the use of a contextualized language model.

One issue that can arise from the word association task is 'out of vocabulary' words. Out of vocabulary words occur when a word either is not contained in the training corpus or when it is so low frequency that it is not retained for the model. If a word is not in the vocabulary of a word embedding model, certain examples will ultimately be omitted from the analysis, or human input will be required. While LSA may be most appropriate for single word comparisons, it is nevertheless the least well suited to handle rare words as it is trained on the smallest and most restrained training corpus. Not only were out of vocabulary words discovered in this task, but so too were references pertaining to the current state of politics and technology like *palmtop* and *gateway 2000* for the source word *computer*, and *Clinton* and *Newt Gingrich* for the source word *politics*. Perhaps the most remarkable is the fact that the word *internet* is not even contained in the original LSA model! Since the model was trained on a corpus that was built before the 21st century, it has the least up to date knowledge of cultural references as compared to models built on newer corpora. Knowledge of the current climate of the world may be an important factor in a task such as word associations, however most word embedding models are not updated over time. This is ultimately an issue that *must* be addressed by first ensuring that an appropriate training corpus is used and then once one is being used, updating the corpus as all embedding models can become outdated.

Another task analyzed for this comparison was where the participants were asked to recite the classic story of Cinderella. For this question, we first sought to understand how well human ratings of content, coherence and tangentiality were predictors of diagnostic class. Here, the ratings of coherence were significant predictors, however content and tangentiality were not. These assays were then operationalized with NLP techniques as an endeavor to understand whether automatization of the ratings could be more sensitive to subtle disruptions in language. The responses were scored with variations of coherence, defined by the average, minimum, maximum and standard deviation of the cosine distances between consecutive windows of words through the recitations of Cinderella. Each approach was computed with LSA, word2vec, GloVe, USE and BERT embeddings with window sizes of 2–8. As there is a wide range of features computed with each approach, the mean, standard deviation, and maximum, as well as the number of significant ($p < 0.01$) F statistics for differentiating patients from controls for each class are given in Table 3. Overall, the features operationalized with BERT and USE generated more significant

Table 3
Comparison of F statistics computed on the minimum, maximum, average, and standard deviation of cosine distances between consecutive window sizes 2–8 (i.e., coherence) through the Cinderella responses with LSA, word2vec, GloVe, USE, and BERT embeddings.

Embedding technique	Mean F statistic	Standard deviation (stdev) F statistic	Maximum F statistic	Significant features (p value < 0.01)
LSA	F = 1.03, p = 0.43	stdev = 1.02	F = 3.43, p = 0.07	N = 0/28
word2vec	F = 1.81, p = 0.35	stdev = 1.83	F = 6.73, p = 0.013	N = 0/28
GloVe	F = 1.90, p = 0.36	stdev = 2.05	F = 6.24, p = 0.02	N = 0/28
USE	F = 2.26, p = 0.35	stdev = 3.08	F = 14.47, p = 0.0004	N = 2/28
BERT	F = 3.41, p = 0.26	stdev = 4.57	F = 18.98, p = 6.9e-05	*N* = 5/28

differentiations between patients and controls. It should be noted though that in most published prediction models, coherence is one feature of several in the model (e.g., Bedi et al., 2015; Elvevåg et al., 2007; Iter et al., 2018) and so may add to the explained variance, but may not be significant as a sole predictor. The overall implication is that while the non-contextualized embeddings may measure the phenomena of interest, the contextualized embeddings generate a more fine-grained understanding of the context of what is being said and, as such, will stand as an appropriate approach for harnessing context in measuring disruptions in language.

Of note is that by using traditional window-based techniques with a powerful tool like BERT or USE, some power may be lost along the way. This is to say that much of the power of techniques like BERT comes from analyzing full paragraphs or sentences with context, and thus when one limits these analyses to comparisons between individual windows of size 2–8, much of their capabilities are also limited. The use of a tool like BERT warrants a *different* approach to calculating disturbances in language. Furthermore, most state-of-the-art successes with BERT are due to the use of large datasets used to fine tune the entire model for predictions, rather than using the pre-trained model to simply extract embeddings for a cosine comparison feature. As such, future work should explore using larger datasets to fine tune BERT to learn features like coherence and tangentiality as a full model with a classifier or regressor as the output layer.

Thus, when designing NLP-based clinical assessments, it is critical to align the assumptions underlying the computed language metrics with the desired goals of the task. For example, we must consider why one embedding model's semantic space or training algorithm would make it well suited for individual clinical tasks. Examples of such considerations would entail determining whether the semantic space appropriately represents the content spoken in patient responses, whether the features computed in the analysis reflect the task constructs (e.g., does it capture the language inherent to the inferencing that is required in the cookie theft picture test), whether it is appropriately sensitive to detect subtle differences in language and semantic changes (e.g., *cookie jar* versus *glass container*).

3.4. Bias in NLP

Bias in AI is a phenomenon in which models may generate prejudiced output in certain cases due to the conscious or unconscious assumptions made during their creation. It is an issue that is unfortunately widespread within the field of AI, however the genre of language-based models introduces additional nuance and vulnerability. This is due to widespread speaker demographic differences and the largely incorrect assumption that users will adhere to standard norms of each language they are speaking, as well as the inevitable identifiable content contained within speech that may cause latent features to be matched to variables of interest. The issue with biased data is that models will tend to fit to the dominant characteristics of the dataset and ignore minority trends. It becomes an issue when this negatively impacts those who exhibit these minority characteristics when the models are unable to perform adequately on their speech and language.

Bias inevitably occurs in nearly every step of the natural language processing pipeline. For instance, this could be in the choice of the data to train models. The majority of studies - especially within psychology - recruit human participants from Western, educated, industrialized, rich and democratic societies (the so-called WEIRD phenomenon; Henrich et al., 2010), and while additional minority classes may be in this group as well, characteristics of a majority class will always overpower the less-represented groups. Further, many off-the-shelf or pretrained language models in particular have been trained on newspaper corpora (Google News, Wikipedia) which are unrepresentative of the manner in which many people speak. Indeed, the median sentence length for spontaneous speech was six words for males and five words for females, where it is more typically 10–15 words in length for technical writing (e.g., Wiggers and Rothkrantz, 2007). Thus, the language and opinions of journalists will inevitably be different from that of a study participant speaking a dialect of English from the American South, for example, and therefore these models may erroneously rate this different language as less coherent.

Bias may also be introduced in the annotation process: if annotators are not familiar with the nuances and norms of the language data that they are annotating, their unfamiliarity may seep into the knowledge that a model synthesizes in the training process. A reported case of this occurring is that of African American English tweets being rated as more toxic simply due to the vernacular used rather than actual toxicity (Sap et al., 2019). This was due to the unfamiliarity of the annotators with their particular language usage. Issues in annotation may additionally arise from the opinions of those who create annotation rubrics and their implicit biases. Next, the choice of representation of the data may introduce a level of bias in that different embedding models are trained on different corpora with different training algorithms and therefore may contain more bias than others. Further, the use of raw speech and language as input into a model may contain more bias-inducing latent features than derived features. Finally, the choice of model used for prediction may affect the level of bias in outputs. For instance, deep learning-based, highly black box models may learn complex and nuanced features that align with demographics in a manner that would be less likely to occur in a simple, traditional machine learning model.

There are known cases of bias particularly in word embeddings. Caliskan et al. (2017) showed that embeddings for traditionally African American names are closer to unpleasant words than pleasant words than traditionally European names. Further, Bolukbasi et al. (2016) showed that the embedding for *she* is closer to the words *homemaker*, *nurse* and *receptionist*, and the embedding for *he* is closer to words *philosopher*, *captain* and *architect*. These issues arise from the distributional statistics of these words showing up in different contexts. As a concrete example, this is due to the fact that language along the lines of “she takes care of the children” and “he works with computers” are more

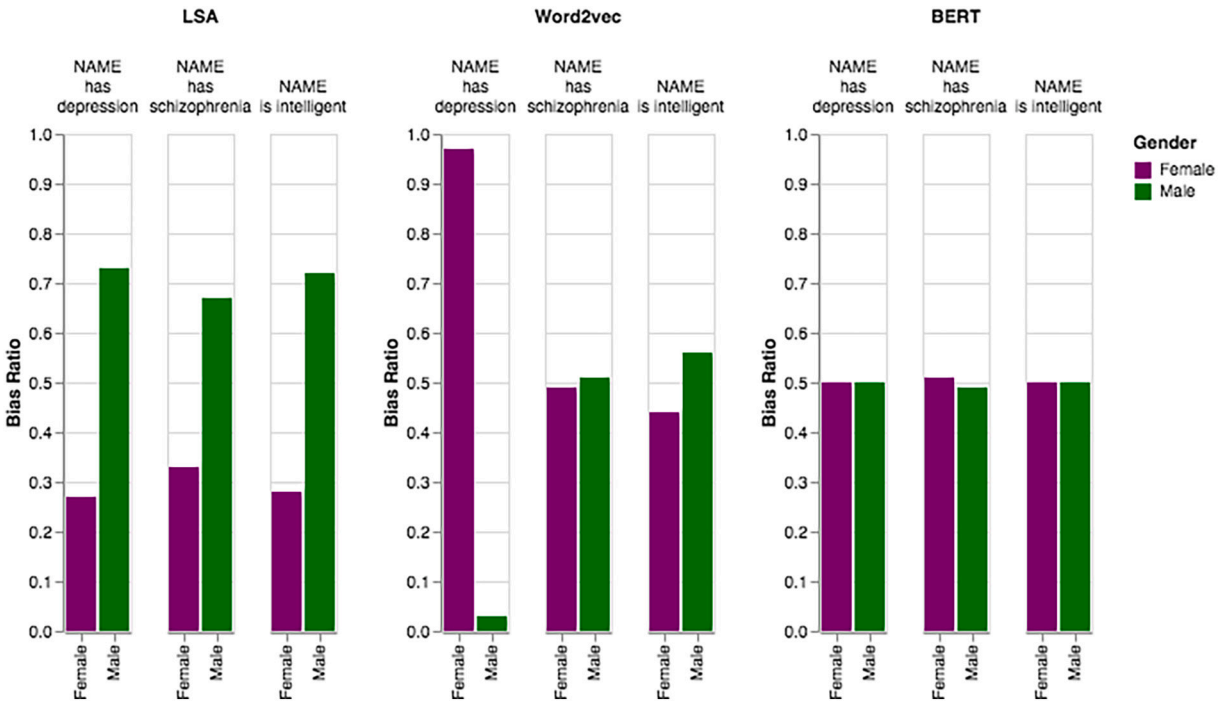

Fig. 2. Ratio of gender bias in LSA, word2vec, and BERT-based cosine distances between the female name Mary/male name James and the phrases "has depression", "has schizophrenia", and "is intelligent". Here, we give a bias measure which computes the proportion of how much the cosine is greater in one gender than the other with a ratio of 0.5 indicating no difference between the two genders.

likely to show up in training corpora than the opposite gender cases. These are cases of race and gender bias that occur in word embeddings simply due to the data the embedding models were trained on. For a review of these issues see, Hovy and Prabhumoye (2021). Reducing bias within the pipeline of machine learning model creation is critical and various approaches have been proposed for doing so. For instance, Bolukbasi et al. (2016) proposed a post hoc method for directly debiasing word embeddings. This method involves separating the embedding spaces by whether the words are gender-specific or gender-neutral and performing linear algebra operations to project the embeddings onto a non-gender specific axis or by making them equidistant to various gender words. Depending on the application of word embedding comparisons, this approach might be critical in the clinical domain (see critical gender bias example in Fig. 2). Beyond computational approaches to minimize bias in language tools, bias-aware considerations should be taken at every step of the application pipeline. At the data collection step, researchers must be thoughtful with respect to balancing important demographic groups and collecting a thorough representation of each group. A common and simple approach to balancing data is to downsample majority classes (i.e., only use a subset of the majority class for training a model so as to balance with a minority sample) or collect more data or create synthetic data to supplement a minority or low resource classes. Overall, what is important is testing for and mitigating bias at every step of the machine learning pipeline: from data collection and annotation to model choice and evaluation. Next, we show how all of these issues - particularly with respect to gender bias - may affect downstream features or predictions in the case of language coherence and disease diagnosis.

3.5. Gender bias

We have described various places in the NLP pipeline where bias may be introduced. Unfortunately, similar sources of bias also exist in research on serious mental illness as it is often the case that male participants outnumber female participants (e.g., Longenecker et al., 2010), which likely negatively impacts subsequent diagnosis and treatment in women. Even in a clinical setting there is much evidence that gender biases the actual diagnosis process. Indeed a 'simulation' study in which clinicians were given the fictional transcripts of clinical scenarios in which the fictional patient was assigned either a male or a female name resulted in a distinct diagnosis bias: clinical scenarios with male names were disproportionately more likely to be diagnosed as having chronic schizophrenia and those with female names tended to be diagnosed as having depression (Høye et al., 2006). Importantly, the clinical scenarios were identical, it was simply the name that was either 'female' or 'male'.

Previous research has not fully established to what extent language derived metrics are gender specific, yet there is already evidence of bias towards men in large corpora (e.g., Bailey et al., 2022). Similarly, there is emerging possible undesirable gender discrimination in clinical applications of AI because of the databases that are leveraged with machine learning (Cirillo et al., 2020; Obermeyer et al., 2019; Sun et al., 2019). By way of illustration, we created the following simple sentences in order to 'compare' how comparisons within three semantic embedding spaces (LSA, word2vec and BERT) changed simply as a function of the putative gender of the character in the sentence: "*NAME* has depression", "*NAME* has schizophrenia", and "*NAME* is intelligent" (where "*NAME*" is substituted by either a female name (Mary) or a male name (James)). Fig. 2 shows the results of these cosine comparisons and uncovers that there are sometimes differences in gender-based similarity to different clinical (or non-clinical) entities. There are also differences between methods employed, where LSA and word2vec display more gender-based differences than BERT. Although these examples are by no means intended to establish that there is gender bias per se in current methods, they serve to illustrate two things, namely that similarity ratings to male and female names can greatly differ and subtle gender bias exists. Based upon the aforementioned human bias in clinical diagnosis we might expect this to be visible also in terms of NLP derived metrics and indeed, may be a reflection of inherent bias in general use of language that became part of the training data for these metrics (Basta et al., 2019). Thus, in use of these NLP measures, we need to ensure that

we have sufficient representation in the training data as well as to evaluate the levels of performance widely in order to minimize the risk of systematic bias.

To sum up the question of *how are we measuring*, we have shown that there are a number of considerations that need to be made by the designer in order to ensure effective, generalizable and unbiased methods. In addition, these design decisions should be explicit and need to be explained to end users of the model. The Data Nutrition Label project (Holland et al., 2018) directly advocates for researchers to be explicit in reporting the distributions of their data in a standardized manner and as such forces researchers to consider what their data entails and potentially uncover potential discrepancies in demographic coverage. Thus, one could envision the scenario where - similar to the mandatory nutrition facts on a cereal box - the 'content' details are listed such as what the training data was composed of (e.g., dataset size, racial makeup), details regarding the model development (e.g., algorithm type), information on its performance (e.g., false positives, false negatives), details about how its assessments (e.g., fairness, bias attestations) and lists of the validation studies (e.g., safety, efficacy). Additionally, it would be important that the 'fact label' also specified what the algorithm's purpose was (e.g., specific illness detection) and critically when the algorithm had last been updated.

4. Why are we measuring?

> *"When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be."*
>
> attributed to Lord Kelvin (1883)

The degree of success in the field of machine learning is often driven by demonstrating high model accuracy, using such measures as agreement to clinician ratings or to predicting diagnostic categories. However, the process of building novel language-based tasks and creating innovative NLP-based features in order to create more accurate models should not be the ultimate goal. Indeed, success should be driven by how the technology and assessments provide *actionable results* that result in improved patient outcomes or improved healthcare delivery. Such outcomes could include more timely detection or lessened workload for clinicians, not requiring travel to a clinic for an in-person visit for patients, improved treatment of a disorder or better quality of life. Actionable results could also include improving our theoretical understanding of the nature of the brain, of patients and/or of disorders. Thus, the creation of any automated assessment should be accompanied with introspection on the purpose and goals of the automation.

4.1. What are we automating?

We can examine the goals of automation in clinical assessment at three levels:

At the first level, we can consider that technology can *automate what humans* do. In this case, the goal of the technology is to take over an existing process that is being done by a clinician and matching the performance of the clinician. This could be, for example, administering and scoring an existing neuropsychological assessment. While this may free up time for a clinician, it is not a highly compelling goal.

The second level of automation is *enhancing what humans do*. For example, automation can perform an initial triage of patients and could work concurrently with, but under the supervision of clinicians. This level can provide levels of efficiency, but can also be able to help clinicians in detecting important cases, where to focus care and when follow-up is needed.

Finally, at the third level, automation can *change or disrupt the existing processes*. Disrupting the process would mean technology changing how the clinical practice is currently done, changing how and when clinicians interact. For example, automation can detect small changes in patient mental states over time to track trajectories and provide individualization to characterize individual baselines rather than population baselines. Such approaches can provide dynamic adaptive testing and improved remote monitoring while *keeping clinicians in the loop*. This further necessitates that the technology provides levels of explainability, trust and support of clinician control so that a clinician still is the ultimate decision-maker (Chandler et al., 2020a). Thus, rather than seeing this level of automation as replacing a clinician, instead, the approach can give stakeholders (e.g., patients, caregivers, clinicians) *more agency on how and when healthcare is applied and on delivering more personalized medicine*. While we should not blindly embrace the latest technological metaphor, if we see that it can bring improvements, we should leverage this but seek to incorporate the best use of both clinicians and computers.

However, automation of clinical decision-making should not be the only goal of applying NLP for mental health. Indeed, there are broad uses for research that go beyond "automating the clinician" and instead provide insights that would not be possible without computational models. We further see that NLP can serve as a forcing function to iteratively improve the construct itself and its psychometric properties, not just the operational definition (e.g., Cohen et al., 2022). For instance, for a construct such as disorganized speech to be computationally instantiated, it requires operationally defining to what degree disorganization is being measured by such features as semantic distance in the choice of words and phrases, changes in syntactic structure or the narrative structure of a recall. Applying machine learning to these features can further reveal insights about underlying neuropsychological mechanisms. By analyzing large numbers of participants' language, we gain insights on how different features correlate and combine to explain relationships to different diagnostic categories and the neural processing.

5. Conclusions

Thus far, the field of computational analyses of language for clinical assessment has been the "wild west", where there have been a variety of approaches implemented and many successes reported as showing the potential. However, the field has focused more on the accuracy of computational models for scoring assessment tasks and predicting diagnostic categories. We need to change our thinking to be about what tasks, computational features and models align best with our understanding of the targeted neurocognitive functions that we want to assess. The field now has enough information about our successes (and failures) to move towards a more principled approach to operationalize our definitions and standardize design and implementation in order to drive applicability of the methods. We advocate for adopting a process where assessments specifically consider the constructs being assessed, how those constructs relate to patient clinical states and how assessment tasks produce language output that can be analyzed by computational methods aligned to the constructs. This principled approach provides a data and method agnostic framework with inferential processes that supports explainability and generalizability in the use of speech as a digital biomarker for mental and cognitive states. This same framework supports the ability to move towards clinician-centered applications.

Despite impressive scientific findings, significant obstacles remain before these techniques will gain acceptance by patients, caregivers and medical providers. We do believe that systems that leverage these measures have translational potential by analyzing large quantities of data to predict optimal and timely interventions. However *just because something is scientifically viable, does not mean it will translate into practice*. Indeed, for any of the aforementioned NLP-based artificial intelligence algorithms to become part of clinical practice, gaining trust of the critical stakeholders, namely patients and clinicians, will be essential. While

the field has thus far focused on *automating what clinicians do in order to meet their level of judgment,* the focus must be switched to *developing accurate tools that incorporate stakeholder needs, are highly transparent and sufficiently explainable and capable of alerting humans to lack of system knowledge or certainty* (e.g., Chandler et al., 2021b; Chandler et al., 2022). Such a collaboration requires multidisciplinarity and the employment of linguistically and culturally diverse data sets, and research that involves all stakeholders and that they are involved in all stages. This will thus make it possible to leverage the very best of NLP/artificial intelligence methods to analyze speech and enable the automation and scaling such that assessment can be conducted remotely and thereby meet the unmet promise of these methods promoting justice, equity, diversity and inclusiveness in important areas of healthcare.

Role of the funding source

The funding source had no role in this publication.

CRediT authorship contribution statement

PWF, CC and BE wrote the first draft of the manuscript. All authors have contributed to the writing of the manuscript, and all have approved submission of this manuscript.

Declaration of competing interest

The authors do not have any conflicts of interests to disclose.

Acknowledgement

Some of this work was supported by a grant from Helse Nord (PFP1301-16) awarded to Brita Elvevåg.

References

Bachman, L.F., 2005. Building and supporting a case for test use. Lang. Assess. Q. 2 (1), 1–34.

Bachman, L.F., Palmer, A.S., 2010. Language Assessment in Practice. Oxford University Press, Oxford.

Bailey, A.H., Williams, A., Cimpian, A., 2022. Based on billions of words on the internet, people = men. Apr Sci Adv. 8 (13), eabm2463. https://doi.org/10.1126/sciadv.abm2463.

Basta, C., Costa-Jussà, M.R., Casas, N., 2019. Evaluating the underlying gender bias in contextualized word embeddings arXiv preprint arXiv:1904.08783.

Bedi, G., Carrillo, F., Cecchi, G.A., Slezak, D.F., Sigman, M., Mota, N.B., Ribeiro, S., Javitt, D.C., Copelli, M., Corcoran, C.M., 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ Schizophr. 1 (1), 1–7.

Bleuler, E., 1911. Dementia Praecox Oder Die Gruppe der Schizophrenien. Deuticke, Leipzig, Germany.

Bleuler, E., 1950. Dementia Praecox or the Group of Schizophrenias. Zinkin J, Translator. International University Press, New York, NY.

Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., Kalai, A., 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Caliskan, A., Bryson, J.J., Narayanan, A., 2017. Semantics derived automatically from language corpora contain human-like biases. Science 356 (6334).

Chaspari, T., Mower Provost, E., Narayanan, S.S., 2013. Analyzing the structure of parent-moderated narratives from children with ASD using an entity-based approach. Proc. Interspeech 2430–2434. Lyon, France.

Cer, D., Yang, Y., Kong, S.Y., Hua, N., Limtiaco, N., John, R.S., Kurzweil, R., 2018. Universal sentence encoder arXiv preprint arXiv:1803.11175.

Chandler, C., Foltz, P.W., Cheng, J., Bernstein, J.C., Rosenfeld, E.P., Cohen, A.S., Holmlund, T.B., Elvevåg, B., 2019. Overcoming the bottleneck in traditional assessments of verbal memory: modeling human ratings and classifying clinical group membership. In: Niederhoffer, K., Hollingshead, K., Resnik, P., Resnik, R., Loveys, K. (Eds.), Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology. Minneapolis, Minnesota, USA, June, pp. 137–147. https://doi.org/10.18653/v1/W19-3016.

Chandler, C., Foltz, P.W., Elvevåg, B., 2020a. Using machine learning in psychiatry: the need to establish a framework that nurtures trustworthiness. Schizophr. Bull. 46, 11–14. https://doi.org/10.1093/schbul/sbz105.

Chandler, C., Foltz, P.W., Cohen, A.S., Holmlund, T.B., Cheng, J., Bernstein, J.C., Rosenfeld, E.P., Elvevåg, B., 2020b. Machine learning for longitudinal applications of neuropsychological testing. Artif. Intell. Med. 1–2, 100006 https://doi.org/10.1016/j.ibmed.2020.100006.

Chandler, C., Holmlund, T.B., Foltz, P.W., Cohen, A.S., Elvevåg, B., 2021a. Extending the usefulness of the verbal memory test: the promise of machine learning. Psychiatry Res. 297, 113743 https://doi.org/10.1016/j.psychres.2021.113743.

Chandler, C., Foltz, P.W., Cohen, A.S., Holmlund, T.B., Elvevåg, B., 2021b. Safeguarding against spurious AI-based predictions: the case of automated verbal memory assessment. In: Proceedings of the NAACL-HLT 2021 Workshop on Computational Linguistics and Clinical Psychology. https://aclanthology.org/2021.clpsych-1.20/.

Chandler, C., Foltz, P.W., Elvevåg, B., 2022. Improving the Applicability of AI for Psychiatric Applications through Human-in-the-loop Methodologies. Schizophr. Bull. 48, 949–957. https://doi.org/10.1093/schbul/sbac038.

Chapelle, C., Enright, M., Jamieson, J., 2010. Does an argument-based approach to validity make a difference? Educ. Meas. Issues Pract. 29, 3–13. https://doi.org/10.1111/j.1745-3992.2009.00165.x.

Chapelle, C.A., 2012. Validity argument for language assessment: the framework is simple…. Lang. Test. 29 (1), 19–27. https://doi.org/10.1177/0265532211417211.

Cirillo, D., Catuara-Solarz, S., Morey, C., et al., 2020. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. NPJ Digit. Med. 3, 81. https://doi.org/10.1038/s41746-020-0288-5.

Cisek, P., 2019. Resynthesizing behavior through phylogenetic refinement. Atten. Percept. Psychophys. 81 (7), 2265–2287. https://doi.org/10.3758/s13414-019-01760-1.

Clarke, N., Foltz, P.W., Garrard, P., 2020. How to do things with (thousands of) words: computational approaches to discourse analysis in Alzheimer's disease. Cortex 129, 446–463.

Cohen, A.S., Le, T., Fedechko, T., Elvevåg, B., 2017. How RDoC can help order thought disorder: the role of psycholinguistics, computational sciences and technology. Schizophr. Bull. 43, 503–508. https://doi.org/10.1093/schbul/sbx030.

Cohen, A.S., Fedechko, T.L., Schwartz, E.K., Le, T.P., Foltz, P.W., Bernstein, J., , et al. Elvevåg, B., 2019. Ambulatory vocal acoustics, temporal dynamics, and serious mental illness. J. Abnorm. Psychol. 128 (2), 97. https://doi.org/10.1037/abn0000397.

Cohen, A.S., Cox, C.R., Masucci, M., Le, T.P., Cowan, T.M., Coghill, L., Holmlund, T.B., Elvevåg, B., 2020a. Digital phenotyping using multimodal data. Curr. Behav. Neurosci. Rep. https://doi.org/10.1007/s40473-020-00215-4.

Cohen, A.S., Schwartz, E., Le, T., Cowan, T., Cox, C., Tucker, R., Foltz, P.W., Holmlund, T.B., Elvevåg, B., 2020b. Validating digital phenotyping technologies for clinical use: the critical importance of "resolution". World Psychiatry 19, 114–115. https://doi.org/10.1002/wps.20703.

Cohen, A.S., Cox, C., Tucker, R., Mitchell, K., Schwartz, E., Le, T., Foltz, P.W., Holmlund, T.B., Elvevåg, B., 2021. Validating biobehavioral technologies for use in clinical psychiatry. Front. Psychiatry 12, 880. https://doi.org/10.3389/fpsyt.2021.503323.

Cohen, A.S., Rodriguez, Z., Warren, K.K., Cowan, T., Masucci, M.D., Edvard Granrud, O., Holmlund, T.B., Chandler, C., Foltz, P.W., Strauss, G.P., 2022. Natural Language Processing and Psychosis: On the Need for Comprehensive Psychometric Evaluation. Schizophr. Bull. 48, 939–948. https://doi.org/10.1093/schbul/sbac051.

Coppersmith, G., Dredze, M., Harman, C., 2014, June. Quantifying mental health signals in Twitter. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pp. 51–60.

Corcoran, C.M., Cecchi, G.A., 2020. Using language processing and speech analysis for the identification of psychosis and other disorders. Biol. Psychiatry: Cogn. Neurosci. Neuroimaging 5 (8), 770–779.

Corcoran, C.M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D.C., et al., 2018. Prediction of psychosis across protocols and risk cohorts using automated language analysis. World Psychiatry. https://doi.org/10.1002/wps.20491.

Cowan, T., Le, T.P., Elvevåg, B., Foltz, P.W., Tucker, R.P., Holmlund, T.B., Cohen, A.S., 2019. Comparing static and dynamic predictors of risk for hostility in serious mental illness: preliminary findings. Schizophr. Res. 204, 432–433. https://doi.org/10.1016/j.schres.2018.08.030.

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., Kumar, M., 2016, May. Discovering shifts to suicidal ideation from mental health content in social media. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 2098–2110.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.

Elvevåg, B., Weinberger, D.R., 2009. Introduction: genes, cognition and neuropsychiatry. Cogn. Neuropsychiatry 14, 261–276.

Elvevåg, B., Foltz, P.W., Weinberger, D.R., Goldberg, T.E., 2007. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. Schizophr Res. 93 (1-3), 304–316. https://doi.org/10.1016/j.schres.2007.03.001.

Elvevåg, B., Foltz, P.F., Rosenstein, M., DeLisi, L., 2010. An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. J. Neurolinguistics 23, 270–284.

Elvevåg, B., Cohen, A.S., Wolters, M.K., Whalley, H.C., Gountouna, V.E., Kuznetsova, K.A., Watson, A.R., Nicodemus, K.K., 2016. An examination of the language construct in NIMH's research domain criteria: time for reconceptualisation! Am. J. Med. Genet. B Neuropsychiatr. Genet. 6 (171), 909–919. https://doi.org/10.1002/ajmg.b.32438.

Elvevåg, B., Foltz, P.W., Rosenstein, M., Ferrer-I-Cancho, R., De Deyne, S., Mizraji, E., Cohen, A., 2017. Thoughts about disordered thinking: measuring and quantifying

the laws of order and disorder. Schizophr Bull. 43 (3), 509–513. https://doi.org/10.1093/schbul/sbx040.
Eyigoz, E., Mathur, S., Santamaria, M., Cecchi, G., Naylor, M., 2020. Linguistic markers predict onset of Alzheimer's disease. EClinicalMedicine 28, 100583. https://doi.org/10.1016/j.eclinm.2020.100583.
Foltz, P.W., 1996. Latent semantic analysis for text-based research. Behav. Res. Methods Instrum. Comput. 28 (2), 197–202.
Foltz, P.W., Laham, D., Landauer, T.K., 1999. The Intelligent Essay Assessor: Applications to Educational Technology. Interact. Multimedia Educ. J. Comput. Enhanc. Learn. 1 (2).
Fraser, K.C., Meltzer, J.A., Rudzicz, F., 2015. Linguistic features identify Alzheimer's disease in narrative speech. J. Alzheimers Dis. 49, 407–422.
Ghanavati, E., Salehinejad, M.A., Nejati, V., et al., 2019. Differential role of prefrontal, temporal and parietal cortices in verbal and figural fluency: implications for the supramodal contribution of executive functions. Sci. Rep. 9, 3700. https://doi.org/10.1038/s41598-019-40273-7.
Goharian, N., Resnik, P., Yates, A., Ireland, M., Niederhoffer, K., Resnik, R., 2021, June. Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Acces.
Guntuku, S.C., Yaden, D.B., Kern, M.L., Ungar, L.H., Eichstaedt, J.C., 2017. Detecting depression and mental illness on social media: an integrative review. Curr. Opin. Behav. Sci. 18, 43–49.
Hawkins, M., Elsworth, G.R., Nolte, S., Osborne, R.H., 2021. Validity arguments for patient-reported outcomes: justifying the intended interpretation and use of data. J. Patient Rep. Outcomes 5 (1), 64. https://doi.org/10.1186/s41687-021-00332-y.
Heinrichs, R.W., Zakzanis, K.K., 1998. Neurocognitive deficit in schizophrenia: a quantitative review of the evidence. Neuropsychology 12 (3), 426.
Henrich, J., Heine, S.J., Norenzayan, A., 2010. The weirdest people in the world? discussion 83-135 Behav. Brain Sci. 33 (2-3), 61–83. https://doi.org/10.1017/S0140525X0999152X.
Hitczenko, K., Cowan, H.R., Goldrick, M., Mittal, V.A., 2022. Racial and ethnic biases in computational approaches to psychopathology. Schizophr. Bull. 48, 285–288. https://doi.org/10.1093/schbul/sbab131.
Hoffman, R.E., 1987. Computer simulations of neural information processing and the schizophrenia-mania dichotomy, 1Feb Arch Gen Psychiatry 44 (2), 178–188. https://doi.org/10.1001/archpsyc.1987.01800140090014.
Hoffman, R.E., Stopek, S., Andreasen, N.C., 1986. A comparative study of manic vs schizophrenic speech disorganization. Sep Arch. Gen Psychiatry 43 (9), 831–838. https://doi.org/10.1001/archpsyc.1986.01800090017003.
Holland, S., Hosny, A., Newman, S., Joseph, J., Chmielinski, K., 2018. The dataset nutrition label: A framework to drive higher data quality standards arXiv preprint arXiv:1805.03677.
Holmlund, T.B., Chandler, C., Foltz, P.W., Cohen, A.S., Cheng, J., Bernstein, J.C., Rosenfeld, E.P., Elvevåg, B., 2020a. Applying speech technologies to assess verbal memory. npj Digit. Med. 3, 33. https://doi.org/10.1038/s41746-020-0241-7.
Holmlund, T.B., Cheng, J., Foltz, P.W., Cohen, A.S., Elvevåg, B., 2019a. Updating verbal fluency analysis for the 21st century: applications for psychiatry. Psychiatry Res. 273, 767–769. https://doi.org/10.1016/j.psychres.2019.02.014.
Holmlund, T.B., Diaz-Asper, C., Elvevåg, B., 2021. The reality of doing things with (thousands of) words in applied research and clinical settings: A commentary on Clarke et al. (2020). Cortex 136, 150–156. https://doi.org/10.1016/j.cortex.2020.08.024.
Holmlund, T.B., Fedechko, T.L., Elvevåg, B., Cohen, A.S., 2020b. Tracking language in real time in psychosis. In: A Clinical Introduction to Psychosis. Academic Press, pp. 663–685. https://doi.org/10.1016/B978-0-12-815012-2.00028-6.
Holmlund, T.B., Foltz, P.W., Cohen, A.S., Johansen, H.D., Sigurdsen, R., Fugelli, P., Bergsager, D., Cheng, J., Bernstein, J., Rosenfeld, E., Elvevåg, B., 2019b. Moving psychological assessment out of the controlled laboratory setting and into the hands of the individual: practical challenges. Psychol. Assess. 31 (3), 292–303. https://doi.org/10.1037/pas0000647.
Hovy, D., Prabhumoye, S., 2021. Five sources of bias in natural language processing. Lang. Linguist. Compass. 1, e12432 https://doi.org/10.1111/lnc3.12432wileyonlinelibrary.com/journal/lnc3.
Høye, A., Rezvy, G., Hansen, V., Olstad, R., 2006. The effect of gender in diagnosing early schizophrenia–an experimental case simulation study. Soc Psychiatry Psychiatr Epidemiol. 41 (7), 549–555. https://doi.org/10.1007/s00127-006-0066-y.
Iter, D., Yoon, J., Jurafsky, D., 2018. Automatic detection of incoherent speech for diagnosing schizophrenia. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pp. 136–146. https://doi.org/10.18653/v1/W18-0615.
Jurafsky, D., Martin, J.H., 2008. Speech and Language Processing: An Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing. Prentice Hall, Upper Saddle River, NJ.
Kane, M.T., 1992. An argument-based approach to validity. Psychol. Bull. 112 (3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527.
Kelvin, W.T., 1883. Popular Lectures and Addresses. Electrical Units of Measurement. 1 (Delivered 3 May 1883).
Landauer, T.K., Foltz, P.W., Laham, D., 1998. Introduction to latent semantic analysis. Discourse Process. 25, 259–284.
Le TP, Moscardini E., Cowan, T., Elvevåg, B., Holmlund, T.B., Foltz, P.W., Tucker, R.P., Schwartz, E.K., Cohen, A.S., 2021. Predicting self-injurious thoughts in daily life using ambulatory assessment of state cognition. J. Psychiatr. Res. 138, 335–341. https://doi.org/10.1016/j.jpsychires.2021.04.013.
Levelt, W.J.M., 1989. Speaking: From Intention to Articulation. MIT Press, Cambridge, MA.
Lezak, M.D., Howieson, D.B., Bigler, E.D., Tranel, D., 2012. Neuropsychological assessment, 5th ed. Oxford University Press.
Longenecker, J., Genderson, J., Dickinson, D., Malley, J., Elvevåg, B., Weinberger, D.R., Gold, J., 2010. Where have all the women gone?: participant gender in epidemiological and non-epidemiological research of schizophrenia. Jun Schizophr Res. 119 (1-3), 240–245. https://doi.org/10.1016/j.schres.2010.03.023.
Low, D.M., Bentley, K.H., Ghosh, S.S., 2020. Automated assessment of psychiatric disorders using speech: a systematic review. Laryngoscope Investig. Otolaryngol. 5 (1), 96–116.
Martin, J.B., 2002 May. The integration of neurology, psychiatry, and neuroscience in the 21st century. Am. J. Psychiatry 159 (5), 695–704. https://doi.org/10.1176/appi.ajp.159.5.695.
Messick, S., 1994. The Interplay of Evidence and Consequences in the Validation of Performance Assessments. Educational Researcher 23 (2), 13–23. https://doi.org/10.2307/1176219.
Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient Estimation of Word Representations in Vector Space. ArXiv:1301.3781 [Cs]. http://arxiv.org/abs/1301.3781.
Mislevy, R.J., Almond, R.G., Lukas, J.F., 2003. A brief introduction to evidence-centered design. ETS Res. Rep. Ser. 2003 (1), i–29.
Morgan, S.E., Diederen, K., Vértes, P.E., SHY, Ip, Wang, B., Thompson, B., Demjaha, A., De Micheli, A., Oliver, D., Liakata, M., Fusar-Poli, P., Spencer, T.J., McGuire, P., 2021. Natural Language Processing markers in first episode psychosis and people at clinical high-risk. Transl Psychiatry. 11 (1), 630. https://doi.org/10.1038/s41398-021-01722-y.
Nicodemus, K.K., Elvevåg, B., Foltz, P.W., Rosenstein, M., Diaz-Asper, C., Weinberger, D. R., 2014. Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach. Cortex 55, 182–191.
Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S., 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366 (6464), 447–453.
Oh, S.J., Sung, J.E., Choi, S.J., Jeong, J.H., 2019. Clustering and switching patterns in semantic fluency and their relationship to working memory in mild cognitive impairment. Dement. Cogn. Disord. 18 (2), 47–61. https://doi.org/10.12779/dnd.2019.18.2.47.
Pennington, J., Socher, R., Manning, C.D., 2014. GloVe: Global Vectors for Word Representation. https://nlp.stanford.edu/pubs/glove.pdf.
Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations. Dec 13.
Prud'hommeaux, E., Roark, B., 2015. Graph-based word alignment for clinical language evaluation. Comput. Linguist. 41 (4), 549–578.
Ranjan, T., Melcher, J., Keshavan, M., Smith, M., Torous, J., 2022. Longitudinal symptom changes and association with home time in people with schizophrenia: an observational digital phenotyping study. Schizophr. Res. 243, 64–69.
Rezaii, N., Walker, E., Wolff, P., 2019. A machine learning approach to predicting psychosis using semantic density and latent content analysis. NPJ Schizophr. 95, 1–12. https://doi.org/10.1038/s41537-019-0077-9.
Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K., Kaye, K., 2011. Spoken language derived measures for detecting mild cognitive impairment. IEEE Trans. Audio Speech Lang. Process. 19 (7), 2081–2090.
Rosenstein, M., Diaz-Asper, C., Foltz, P.W., Elvevåg, B., 2014. A computational language approach to modeling prose recall in schizophrenia. Cortex 55, 148–166.
Rosenstein, M., Foltz, P.W., DeLisi, L.E., Elvevåg, B., 2015a. Language as a biomarker in those at high-risk for psychosis. Schizophr. Res. 165, 249–250.
Rosenstein, M., Foltz, P.W., Vaskinn, A., Elvevåg, B., 2015b. Practical issues in developing semantic frameworks for the analysis of verbal fluency data: A Norwegian data case study. Denver, Colorado, June 5, 2015. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. Association for Computational Linguistics, pp. 124–133. http://m-mitchell.com/clpsych2015/pdf/CLPsych15.pdf.
Santeramo, R., Withey, S., Montana, G., 2018. Longitudinal detection of radiological abnormalities with time-modulated LSTM. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, Cham, pp. 326–333.
Sap, M., Card, D., Gabriel, S., Choi, Y., Smith, N.A., 2019. The Risk of Racial Bias in Hate Speech Detection. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp. 1668–1678.
Strauss, E., Sherman, E.M.S., Spreen, O., 2006. A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary, 3rd ed. Oxford University Press.
Sun, T., et al., 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp. 1630–1640. https://aclanthology.org/P19-1159.
Tagamets, M.A., Cortes, C.R., Griego, J.A., Elvevåg, B., 2014. Neural correlates of the relationship between discourse coherence and sensory monitoring in schizophrenia. Cortex 55, 77–87.
Troyer, A.K., 2000. Normative data for clustering and switching on verbal fluency tasks. J. Clin. Exp. Neuropsychol. 22 (3), 370–378. https://doi.org/10.1076/1380-3395(200006)22:3;1-V;FT370.
Trull, T.J., Ebner-Priemer, U.W., 2009. Using experience sampling methods/ecological momentary assessment (ESM/EMA) in clinical assessment and clinical research: introduction to the special section. Psychol. Assess. 21 (4), 457–462. https://doi.org/10.1037/a0017653.
Voorspoels, W., Storms, G., Longenecker, J., Verheyen, S., Weinberger, D.R., Elvevåg, B., 2014. Deriving semantic structure from category fluency: clustering techniques and their pitfalls. Cortex 55, 130–147.

Voppel, A.E., de Boer, J.N., Brederoo, S.G., Schnack, H.G., Sommer, I., 2021. Quantified language connectedness in schizophrenia-spectrum disorders. Psychiatry Res. Oct, 304:114130. https://doi.org/10.1016/j.psychres.2021.114130.

Wechsler, D., 2009. Wechsler Memory Scale - Fourth Edition, WMS-IV: Technical and Interpretive Manual. Pearson, San Antonio, TX.

Wiggers, P., Rothkrantz, L.J.M., 2007. Exploratory analysis of word use and sentence length in the spoken Dutch corpus. Jun. In: Matoušek, V., Mautner, P. (Eds.), Text, Speech and Dialogue. TSD 2007. Lecture Notes in Computer Science, 4629. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-74628-7_48.

Yudofsky, S.C., Hales, R.E., 2002. Neuropsychiatry and the future of psychiatry and neurology. Aug Am J Psychiatry. 159 (8), 1261–1264. https://doi.org/10.1176/appi.ajp.159.8.1261.

Zhao, J., Feng, Q., Wu, P., et al., 2019. Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. Sci. Rep. 9, 717. https://doi.org/10.1038/s41598-018-36745-x.