Прогнозирование психоза на основании разных протоколов в группах высокого риска с помощью метода автоматизированного анализа языка

Cheryl M. Corcoran[1,2], Facundo Carrillo[3,4], Diego Fernandez-Slezak[3,4], Gillinder Bedi[2,5,6], Casimir Klim[2,5], Daniel C. Javitt[2,5], Carrie E. Bearden[7], Guillermo A. Cecchi[8]

[1]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA;
[2]New York State Psychiatric Institute, New York, NY, USA;
[3]Departamento de Computacion, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina;
[4]Instituto de Investigacion en Ciencias de la Computacion, Universidad de Buenos Aires, Buenos Aires, Argentina;
[5]Department of Psychiatry, Columbia University Medical Center, New York, NY, USA;
[6]Centre for Youth Mental Health, University of Melbourne, and Orygen National Centre of Excellence in Youth Mental Health, Melbourne, Australia;
[7]Department of Psychiatry and Biobehavioral Sciences and Psychology, University of California Los Angeles; Semel Institute for Neuroscience and Human Behavior, Los Angeles, CA, USA;
[8]Computational Biology Center – Neuroscience, IBM T.J. Watson Research Center, Ossining, NY, USA

Перевод: Дорофейкова М.В. (Санкт-Петербург, Россия)
Редактура: к.м.н. Смирнова Д.А. (Самара, Россия, Перт, Австралия)

Резюме

Язык и речь являются основным источником сведений для психиатров на этапах диагностики и лечения психических расстройств. При психозе может нарушаться сама структура языка, в частности, семантическая связность (например, соскальзывание и резонерство) и синтаксическая сложность (например, конкретность). Трудноуловимые нарушения речи проявляются при шизофрении уже в доманифестный период психоза, на продромальных стадиях болезни. С помощью метода компьютеризированного автоматизированного анализа естественного языка, мы ранее показали, что среди англоговорящих молодых людей с клинически высоким (например, ультравысоким) риском развития психоза изначальное снижение семантической связности (смысловой стройности речи) и синтаксической сложности может с высокой точностью предсказывать последующее наступление психоза. Целью нашей работы были валидация автоматизированных методов лингвистического анализа во второй, более крупной выборке из группы риска, также в англоговорящей популяции, и выявление различий речи при психозе и в норме. Нам удалось выделить автоматизированный, основанный на машинном обучении, классификатор речи, включающий показатели сниженной семантической связности, ее большей вариабельности и уменьшения частоты использования притяжательных местоимений, и прогнозирующий наступление психоза с точностью 83% (по внутреннему протоколу), 79% – при кросс-валидации методики в первой когорте (по перекрестному протоколу) и 72% – при дифференциации речи здоровых и больных с недавней манифестацией психоза. Классификатор продемонстрировал сильную корреляцию с ранее выявленными вручную лингвистическими предикторами. Наши результаты подтверждают полезность и валидность автоматизированных методов обработки естественного языка для оценки особенностей нарушений речи на уровне семантики и синтаксиса на разных стадиях психотического расстройства. Следующие этапы работы будут включать применение этих методов в более крупных группах риска для дальнейшей проверки воспроизводимости полученных результатов, в том числе на других языках, помимо английского, и выявления источников вариабельности. Данная технология может повысить качество прогнозирования исходов психоза у молодых людей группы высокого риска и определить лингвистические мишени для языковой реабилитации и профилактических вмешательств. В целом и общем, автоматизированный лингвистический анализ может стать мощным инструментом диагностики и лечения в нейропсихиатрии.

Ключевые слова: автоматизированный лингвистический анализ, прогнозирование психоза, семантическая связность, синтаксическая сложность, молодые люди из группы высокого риска, машинное обучение.

(World Psychiatry 2018;17(1):67-75)

Язык предоставляет нам возможность уникальным образом заглянуть внутрь психики: на основе речевых высказываний мы делаем заключения о мыслительных процессах других людей таким образом, что, например, считаем дезорганизованный язык отражением имеющихся расстройств мышления. Языковые нарушения широко распространены при шизофрении и непосредственно связаны со снижением функционирования пациентов, учитывая тот очевидный факт, что человек должен четко думать и изъясняться, чтобы поддерживать общение с друзьями и сохранять работу[1]. При шизофрении говорящий «нарушает привычные синтаксические и семантические правила, которые регулируют использование языка» таким образом, что это выражается в снижении синтаксической сложности (конкретная речь, бедность содержания речи) и потере семантической связности, например в форме перерывов смыслового потока в речи (разорванное мышление, резонерство)[2]. Данные языковые нарушения представляют собой ключевую характерную черту шизофрении и клинически выявляются в форме трудноуловимых изменений речи в группах молодых людей с высоким семейным[3] и клиническим[4-7] риском развития психоза уже на ранних этапах заболевания в доманифестный период.

С целью повышения качества прогноза манифестации психоза, помимо клинических шкал, предпринимались попытки охарактеризовать данные рано проявляющиеся и трудноуловимые языковые нарушения у пациентов с клинически высоким риском развития психоза (clinical high-risk, CHR), используя методы линвгистического анализа. Bearden и соавт.[8] исследовали короткие образцы записей речи представителей группы CHR с применением лингвистического анализа с помощью кодирования вручную и обнаружили, что как семантические изменения (нелогичное мышление), так и снижение синтаксической сложности (бедность речи) предсказывали наступление психоза с точностью 71%, по сравнению с точностью 35%, характерной для клинических оценок. Снижение показателя референциальной связности, проявляющееся в таком использования

местоимений и компаративов («это» или «то»), при котором зачастую невозможно прояснить, о ком конкретно или о чем именно из ранее упомянутого идет речь в высказываниях пациента, также предсказывало начало психоза.

Хотя такой лингвистический подход с обработкой речевых данных вручную очевидно предсказывает развитие психоза лучше, чем клинические оценки, он зависит от заранее определяемых мер и критериев измерений, которые могут не охватывать все трудноуловимые языковые особенности. В связи с этим для анализа речи в группах CHR мы использовали автоматизированные методы обработки естественного языка. Данный вид вероятностного лингвистического анализа базируется на сборе компьютером лексики (семантики) и изучении грамматики (синтаксиса) с помощью алгоритмов машинного обучения, выстроенных на основании изучения больших объемов текста, что, в свою очередь, оказалось возможным благодаря экспоненциальному увеличению вычислительной мощности компьютеров и потокам текстов, которые оказались доступными с появлением интернета.

Что касается семантики, то здесь общепринятым подходом является латентно-семантический анализ, который разработан на основании теорий усвоения словарного запаса[9,10] и в рамках которого значение слова исследуется через показатели его встречаемости в сочетании с другими словами. При использовании данного метода каждому слову присваивается многомерный семантический вектор, так, что косинус между словами-векторами отражает семантическое сходство между словами. Группирование расположенных рядом друг с другом слов-векторов может использоваться для оценки семантической связности нарратива.

Исследования речи больных шизофренией с помощью латентно-семантического анализа позволили выявить взаимосвязь между снижением семантической связности, клиническими оценками расстройств мышления и функционального снижения, а также аномальной активацией языковых нейронных сетей в ответ на выполнение заданий[11,12].

Для оценки синтаксиса используется метод выделения (тегирования) фрагментов речи, который позволяет определить длину предложения и частоту использования различных частей речи[13,14].

В более раннем исследовании, ориентированном на подтверждение концепции и проведенном по протоколу с нарративами на небольшой выборке CHR, мы использовали как латентно-семантический анализ, так и выделение фрагментов речи, с помощью метода машинного обучения, для того чтобы выделить классификатор, который включал минимальные значения семантической связности, сокращения длины предложения и снижения частоты употребления определительных местоимений (например, «чтобы» или «который»), используемых для введения придаточных предложений[15]. Эти три характеристики не только коррелировали с клиническими оценками, но даже превзошли их в отношении точности прогнозирования психоза.

В настоящем исследовании мы использовали тот же самый автоматизированный подход для обработки естественного языка с помощью метода машинного обучения, включавший латентно-семантический анализ и выделение фрагментов речи, но на этот раз мы применили его в рамках протокола с расширенной оперативной базой данных речи группы CHR, который ранее анализировали Bearden и соавт.[8] с помощью лингвистических методов с кодированием вручную.

Мы предположили, что классификатор, основанный на изучении расширенного набора данных[8], будет очень точным (около 80%) в прогнозировании наступления психоза как при тестировании в рамках основного протокола, так и при повторном тестировании в рамках протокола с нарративами[15] (перекрестный протокол). Мы также предположили, что лингвистические показатели, выделенные на основании базы данных обучения автоматизированным способом и способом вручную, будут коррелировать между собой.

Далее мы проверили способность классификатора отличать речь подростков с недавно начавшимся психозом от нормальной речи, т. е. тем самым выступать в роли предполагаемого раннего маркера заболевания.

МЕТОДЫ

Участники

Исследование проводилось на базе Калифорнийского университета Лос-Анджелеса (University of California Los Angeles, UCLA) и включало 59 участников с CHR. Участники соответствовали критериям включения в рамках одной из трех категорий продромальных синдромов, которые оценивались с помощью Структурированного интервью для выявления продромальных синдромов/Шкалы продромальных синдромов (Structured Interview for Prodromal Syndromes/Scale of Prodromal Symptoms, SIPS/SOPS)[16]: а) подпороговые позитивные симптомы, б) краткие интермиттирующие психотические симптомы, в) существенное снижение в социальном/ролевом функционировании в сочетании с диагнозом шизотипического расстройства личности или при наличии родственника первой линии родства, страдающего психотическим расстройством. Из числа представителей данной группы у 19 человек в течение двух лет развилось психотическое расстройство («конвертеры», CHR+), у 40 – не выявилось (CHR–). Процесс выхода в психоз оценивался с помощью критериев «наличия психоза» SIPS/SOPS. Помимо этого, исследовались записи 16 пациентов UCLA с недавно развившимся психозом и 21 здорового человека, сопоставимых по демографическим показателям и рекрутированных из местных школ и сообщества.

В группу исследования также вошли 34 участника из Нью Йорка, чей статус CHR с помощью вышеописанных критериев SIPS/SOPS. Из их числа, в соответствии с критериями SIPS/SOPS, у пяти человек в течение 2,5 года развился психоз (CHR+), у 29 психоза выявлено не было (CHR–).

Таблица 1. Демографические характеристики двух групп исследования

	UCLA				Нью-Йорк	
	CHR+ (N=19)	**CHR– (N=40)**	**КГ (N=21)**	**ППЭ (N=16)**	**CHR+ (N=5)**	**CHR– (N=29)**
Возраст на момент начала исследования (годы, ср. знач.±с.о.)	17,3±3,7	16,4±3,0	18,0±2,8	15,8±1,7[a]	22,2±3,4	21,2±3,6
Пол (% мужчин)	89,5	55,0[б]	61,9[б]	68,7	80,0	65,5
Этническая принадлежность (% европеоидов)	63,1	50,0	66,7	62,5	40,0	37,9
Социально-экономический статус родителей (индекс Холлингшеда, ср. знач.±с.о.)	4,4±2,1[a]	4,4±1,7[a]	5,7±1,4	4,9±1,8	Нет данных	Нет данных

*Значимые различия на уровне $p<0,05$: [a]по сравнению с контрольной группой, [б]по сравнению с группой CHR+.
**CHR+ – лица с клинически высоким риском, у которых психоз развился за период наблюдения, CHR– – лица с клинически высоким риском, у которых психоз не развился за период наблюдения, КГ – контрольная группа из числа здоровых людей, ППЭ – пациенты с первым психотическим эпизодом.

Демографические характеристики двух групп представлены в табл. 1. Локальные этические комитеты одобрили проведение исследования, а все участники дали информированное согласие (в случае несовершеннолетних участников – согласие родителей).

Набор речевого материала

UCLA (протокол базы данных с подсказками)

Исследовалась речь, полученная при использовании «Игры в рассказ» Каплана, во время которой участники пересказывали услышанную историю и потом отвечали на вопросы относительно ее содержания («Что Вам нравится в этой истории?», «Правдивая ли это история?»), а впоследствии сочиняли и рассказывали новую историю[17]. Записи речи были транскрибированы и деидентифицированы, т. е. такие соответствующие существительные, как имена собственные, были заменены.

Лингвистический анализ с обработкой данных вручную включал в себя применение рейтинговой шкалы формального расстройства мышления Кидди (Kiddie Formal Thought Disorder Rating Scale – K-FTDS) и использование подхода Halliday и Hassan для анализа связности в модификации Каплана[17]. Баллы K-FTDS включали количественные показатели частоты встречаемости элементов нелогичного мышления, феномена утраты ассоциаций и обеднения содержания речи. Среди категорий, определяющих связность, оценивались референциальные показатели (пронаимальные, указательные и сравнительные – «это», «то»), союзы («и», «но», «потому что») и нечеткости/двусмысленности[17]. Этот набор данных использовался для анализа точности прогнозирования внутри протокола.

NYC (протокол базы данных с нарративами)

Интервью для сбора нарративов продолжительностью около одного часа с отсутствием четкого ограничения по времени проводились интервьюерами, специально подготовленными экспертом по вопросам качественных методов исследования. Подсказки давались с целью направить разговор на темы переживания опыта жизненных изменений и ожиданий относительно будущего[18]. Этот набор данных использовался для изучения точности прогнозирования психоза в перекрестном протоколе.

Анализ речи

Предварительная обработка речи

Стенограммы речи были предварительно обработаны и подготовлены для компьютерного анализа. Мы использовали программу Natural Language Toolkit, свободно доступную в интернете (NLTK; http://www.nltk.org). Вначале были исключены знаки препинания (например, запятые, точки), слова были подразделены на лексемы (идентифицированы как части речи), а затем каждый речевой образец был разобран на фразы с использованием правил грамматики английского языка. Потом слова были конвертированы в корневые формы слов, от которых они изначально образованы, или разделены на леммы с использованием программы лемматизатора NLTK WordNet.

Полученные обработанные речевые данные для каждого речевого образца представляли собой ряды слов, сохраненных в первоначальном порядке их произнесения и подразделенных на леммы, без пунктуации и записанных в нижнем регистре.

Латентно-семантический анализ

Латентно-семантический анализ[9,10] использовался для преобразования содержания каждого речевого образца из ряда слов в ряд семантических векторов при сохранении первоначального порядка изложения текста. При данном типе анализа каждому слову в лексиконе присваивается многомерный семантический вектор на основе показателей его совместной встречаемости в сочетании с другими словами на основании представленности в расширенном текстовом корпусе, в частности, в корпусе коллекции образовательных материалов Touchstone Applied Science Associates (TASA).

Автоматизированный анализ позволяет выстраивать системы смыслов в речи, как бы воссоздавая процессы, происходящие в психике человека, когда значения слов определяются исходя из опыта предыдущего контакта с данными словами в различных контекстах. Компьютер «узнает» значения слов с помощью вычислений, сканируя расширенные тексты и определяя частоту совпадения каждого отдельного слова с любым другим словом в лексиконе. Считается, что слова, которые встречаются вместе чаще, имеют большее семантическое сходство (например, «кошка»/«собака» в отличие от «кошка»/«карандаш»), и направление их векторов будет в большей степени схожим. Сочетания слов (например, предложения) имеют семантические векторы, которые представляют собой суммы семантических векторов всех слов, содержащихся в данном словосочетании (предложении). Семантическая связность между словами или между словосочетаниями (например, последовательными предложениями) может быть вычислена путем подсчета значения косинуса между последовательными семантическими векторами (от -1,0 при несогласованности до 1,0 при связности).

Поскольку протокол с нарративами в Нью-Йорке подразумевал отсутствие ограничения по времени окончания повествования и непрерываемые ответы состояли в среднем из 130 слов для группы CHR– и 182 слов для CHR+, в нашем предыдущем исследовании было получено достаточно образцов свободной речи для анализа семантической связности на уровне предложений[15]. В то же время протокол исследования на базе UCLA, основанный на подсказках[8], позволил получить гораздо более короткие ответы (средняя длина непрерываемого ответа составила менее 20 слов; недостаточное количество предложений для анализа), поэтому для оценки результатов использовалось значение k-уровня семантической связности, которое предполагает вычисление вариабельности k расстояний между словами, где k находится в диапазоне от 5 до 819. Как и в нашем предыдущем исследовании[15], мы рассчитали типовые статистические показатели для каждого из измерений k-уровня связности, такие как среднее, стандартное отклонение, минимальное, максимальное значения и 90-е процентили (менее чувствительны к выбросам, чем максимальные значения), они были также «нормализованы» или скорректированы в зависимости от длины предложения.

Анализ с выделением фрагментов речи

Так же, как каждому слову в отдельном речевом образце был присвоен семантический вектор, каждое слово было помечено в отношении своей грамматической функции с помощью процедуры POS-Tag в программе Natural Language Toolkit (www.nltk.org), находящейся в открытом доступе и разработанной на основании корпуса слов, проанализированных вручную, в рамках базы Penn Treebank13. Например, предложение «The dog is near the fence»/ «Собака находится около забора» приобрело бы следующие пометки (теги): («The»/артикль, «DT»), («dog»/ «собака», «NN»), («is»/ «находится», «VBZ»), («near»/ «около», «IN»), («the»/артикль, «DT»), («fence»/ «забора», «NN»), где пометка DT означает определители, NN – существительные, VBZ – глаголы и IN – предлоги.

Penn Treebank включает набор из 36 тегов фрагментов речи, которые характеризуют типы существительных, глаголов, прилагательных, наречий, определителей, предлогов и местоимений. Для каждого речевого образца мы вычисляли частоту использования каждой грамматической функции.

Классификация на основе машинного обучения

Алгоритм машинного обучения классифицирует речь по тому принципу, содержит ли она в себе характеристики, свойственные лицам, у которых впоследствии развивается психоз, и учитывая характеристики тех, у кого психоз не выявляется. Для того чтобы это осуществить, алгоритм изучает выборки речевых образцов на предмет имеющихся паттернов, а затем итеративным способом предсказывает классификацию (психоз или нет) в новых речевых образцах, не использованных на этапе обучения.

Анализ, основанный на машинном обучении, был ограничен одиннадцатью переменными речи, которые продемонстрировали значимые различия между группами CHR+ и CHR– в выборке UCLA (девять характеристик семантической связности и два синтаксических элемента – частота встречаемости сравнительных форм прилагательных и притяжательных местоимений), а также тремя переменными, которые предсказывали психоз по данным наших предыдущих исследований[15], включая определители, местоимения и прилагательные из «семейства WH» («which»/ «который», «what»/ «что», «whom»/ «кому»). Перечень данных четырнадцати элементов, используемых для анализа, приводится в табл. 2. Каждый речевой образец имел вектор, состоящий из этих четырнадцати переменных.

Для прояснения внутренней структуры и взаимосвязи речевых данных, мы добавили в анализ данные выборки здорового контроля UCLA и сделали дополнительную процедуру разложения векторов речевых образцов на единичные значения по всем четырнадцати языковым показателям (данная процедура представляет собой один из видов факторного анализа, основанного на линейной алгебре). На основании результатов этого анализа мы выбрали четыре ведущих фактора, которые лучше остальных позволяли различать речь лиц с клинически высоким риском психоза CHR+ и речь представителей группы CHR–. Впоследствии на основании этих четырех факторов, используя итерации обучения в отдельной выборке образцов и прогнозирование на основании данных оставшейся выборки, мы построили модель логистической регрессии для разделения групп CHR+ и CHR–.

Перекрестная валидация

Те же самые четырнадцать характеристик были извлечены из данных, собранных в Нью-Йорке, и соотнесены с характеристиками выборки UCLA с использованием простого глобального «прокрустова» преобразования координат[20,21], напоминающего пространственную регистрацию при нейровизуализации[22], которая включает масштабирование (по размеру), вращение и перевод в евклидово пространство. Данный анализ позволил минимизировать разницу в ковариации двух наборов данных, сохраняя при этом относительное взаиморасположение между точками данных.

Кроме того, мы применили встроенный метод алгоритмов построения выпуклой оболочки множеств, использованный в наших предыдущих исследованиях[15], чтобы создать трехмерное пространство (на основании трех первых факторов) для моделирования точности классификатора, полученного на когорте UCLA, для различения CHR+ и CHR– в преобразованной выборке из Нью-Йорка. Выпуклая оболочка множества точек представляет собой минимальный выпуклый многогранник, который их размещает.

Корреляции между текстовыми особенностями, демографическими, клиническими характеристиками и лингвистическими показателями, полученными при кодировании вручную.

Мы оценили наличие взаимосвязи между четырнадцатью выявленными характеристиками речи, показателями возраста, пола, этнической принадлежности (европеоиды/ неевропеоиды) и социально-экономического статуса родителей[23]. В дальнейшем мы проверили, существуют ли корреляционные связи особенностей текста с клиническими показателями или с тремя лингвистическими характеристиками, закодированнными вручную (нелогичность, бедность содержания и референциальная связность), которые предсказывали манифестацию психоза в когорте UCLA в ранее проведенном исследовании[8]. Мы рассчитали каноническую корреляцию между автоматизированными и закодированными вручную текстовыми переменными, которая отражает взаимосвязь между двумя наборами лингвистических переменных, полученных от одних и тех же индивидов.

Полезность классификатора для дифференциации речи при психозе и в норме

В рамках независимой валидации мы определяли точность речевого классификатора CHR при различении речи 21 здорового добровольца и 16 пациентов с недавно развившимся психозом из группы UCLA по протоколу с подсказками. Идея заключалась в том, что представители здорового контроля должны иметь речь, похожую на речь CHR–, в то время как речь пациентов с недавним началом психоза должна быть похожа на речь CHR+.

Таблица 2. Синтаксические и семантические элементы, используемые для прогностического моделирования

Описание	**Пример**
a. Прилагательное, компаратив	«смелее», «ближе», «милее»
b. Притяжательное местоимение	«ее», «его», «мое», «мой», «наш», «наш», «их», «ваш»
c. WH-определитель	«что», «который» – «that», «which», «what»
d. WH-местоимение	«что», «который», «кто», «кого» – «that», «what», «which», «who», «whom»
e. WH-наречие	«как», «однако», «когда-либо», «почему» – «how», «however», «whenever», «why»
f. Минимум показателя связности на 5-м уровне, нормализованный	
g. Минимум показателя связности на 5-м уровне	
h. 90-й процентиль показателя связности на 5-м уровне	
i. Максимум показателя связности на 6-м уровне	
j. Среднее значение показателя связности на 7-м уровне, нормализованное	
k. Стандартное отклонение показателя связности на 7-м уровне, нормализованное	
l. 90-й процентиль на 7-м уровне	
m. Стандартное отклонение показателя связности на 7-м уровне	
n. 90-й процентиль на 8-м уровне	

* Использованы значения показателя k-уровня семантической связности, который отражает вариабельность «k» расстояний между словами, где k находится в диапазоне значений от 5 до 8.

РЕЗУЛЬТАТЫ

Классификация, основанная на машинном обучении

Из четырех факторов в классификаторе, основанном на машинном обучении, первые три отражали взвешенные значения семантических признаков (соответственно, максимум, минимум и разброс значений показателя семантической связности), в то время как четвертый фактор являлся взвешенной оценкой частоты использования притяжательных местоимений (рис. 1).

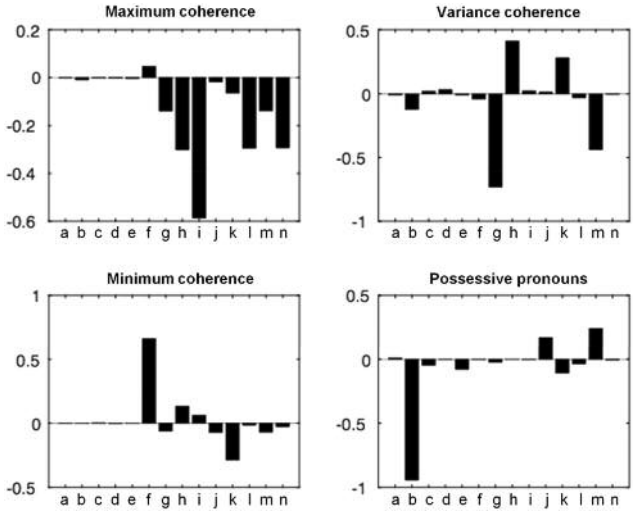

Рис. 1. Основанный на машинном обучении четырехфакторный классификатор для предсказания вероятности развития психоза от Калифорнийского университета Лос-Анджелеса (UCLA). Факторы представляют собой совокупности взвешенных значений синтаксических характеристик (a–e) и семантических характеристик связности (f–n), перечисленных в табл. 2. Первые три фактора взвешены по отношению к семантическим показателям (максимум, разброс значений и минимум), а четвертый фактор взвешен по отношению к синтаксической характеристике (притяжательные местоимения). Оси Y показывают взвешенные коэффициенты

Точность, с которой совокупность этих четырех факторов классифицировала исход психоза в когорте UCLA, составила 83% с использованием классификатора логистической регрессии. Ретроспективный анализ позволил определить площадь под кривой (AUC) равную 0,87 на ROC-кривой (receiver operating characteristic, рабочая характеристика приемника; рис. 2).

Таким образом, классификатор, включавший показатели сниженной семантической связности, большей вариабельности значений связности и сниженной частоты употребления притяжательных местоимений («ее», «его», «мое», «мое», «наше», «наше», «их», «ваше»), оказался очень точным в прогнозировании наступления психоза.

Перекрестная валидация

Результаты использования классификатора UCLA применительно к первоначальному набору речевых данных из Нью-Йорка, после проведения «прокрустова» преобразования координат[20,21,24], показали, что он способен успешно подразделять группу CHR по фактору наступления психоза (p<0,05) со специфичностью 0,82 (24/29) и чувствительностью 0,60 (3/5), так что общая точность составляет 0,79. Анализ преобразованных речевых данных когорты из Нью-Йорка показал, что при использовании логистической регрессии применение классификатора UCLA характеризуется значением AUC 0,72 (см. рис. 2).

С целью сравнения полученных результатов с результатами нашего предыдущего исследования[15] мы создали трехмерную проекцию данных с использованием трех ведущих факторов, выделенных из набора речевых данных UCLA CHR. Мы получили выпуклые оболочки, которые исключили 11 из 19 CHR+ в когорте UCLA (т. е. 8/19 ложноотрицательных результатов; рис. 3A); это указывает на то, что классификатор с применением логистической регрессии (со всеми четырьмя факторами) является более точным. С помощью тех же трех факторов с классификатором UCLA выпуклая оболочка CHR– из Нью-Йорка исключила трех из пяти CHR+ (рис. 3B). Следует отметить, что наблюдалось существенное перекрытие выпуклых оболочек индивидов CHR– для двух наборов речевых данных (из UCLA и Нью-Йорка; рис. 3C).

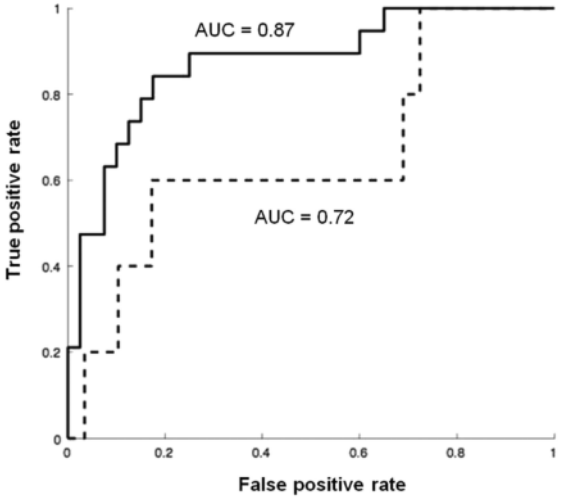

Рис. 2. Рабочая характеристика приемника (ROC) классификатора исхода психоза при клинически высоком риске UCLA, примененного по отношению к группе участников исследования на базе UCLA (сплошная линия) и в Нью-Йорке (пунктирная линия). AUC – площадь под кривой

Корреляции с демографическими характеристиками, клиническими оценками и лингвистическими показателями, полученными при кодировании вручную

Среди демографических характеристик фактор возраста продемонстрировал статистически значимую взаимосвязь с тремя показателями семантической связности: 90% для 5-го уровня (p=0,002), 7-го уровня (p=0,01) и 8-го уровня (p=0,004), что указывает на повышение семантической связности с возрастом. Напротив, связи между текстовыми переменными, полученными с помощью автоматизированного анализа, и факторами пола, этнической принадлежности или социально-экономического статуса родителей выявлено не было[23].

Значимой связи между характеристиками речи, полученными с помощью автоматизированного анализа текста, и показателями клинической оценки по SIPS/SOPS (суммарными позитивной и негативной оценками) также не было выявлено. Однако каноническая корреляция между четырнадцатью выявленными речевыми особенностями и тремя лингвистическими показателями, полученными при кодировании вручную (нелогичность, бедность содержания, реляционная сплоченность), которые предсказывали наступление психоза в предыдущем исследовании[8], характеризовалась большой силой эффекта и высокой статистической значимостью, с r=0,71, p<10^{-6}.

Полезность классификатора для дифференциации речи при психозе и в норме

Применение классификатора с использованием логистической регрессии в отношении набора речевых данных здоровых лиц и пациентов с недавно начавшимся психозом из выборки UCLA позволило дифференцировать речь при психозе и нормальную речь с точностью 72%.

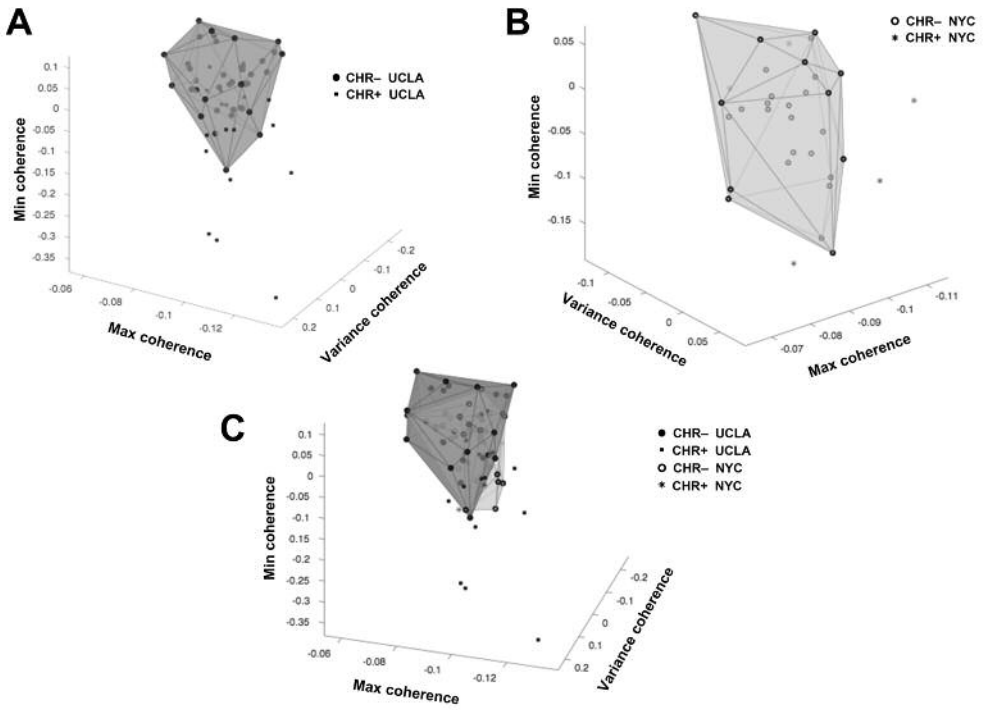

Рис. 3. Проекция трех ведущих факторов для группы CHR из Калифорнийского университета Лос-Анджелеса (UCLA) и из Нью-Йорка (NYC). Данные факторы были взвешены по показателям семантической связности. A. Выпуклая оболочка CHR– в Лос-Анджелесе с 11 из 19 CHR+ за пределами оболочки. B. Выпуклая оболочка CHR– из Нью-Йорка с 3 из 5 CHR+ вне ее. C. Данные из A и B (все CHR) показаны вместе для демонстрации степени перекрытия значений речевых показателей

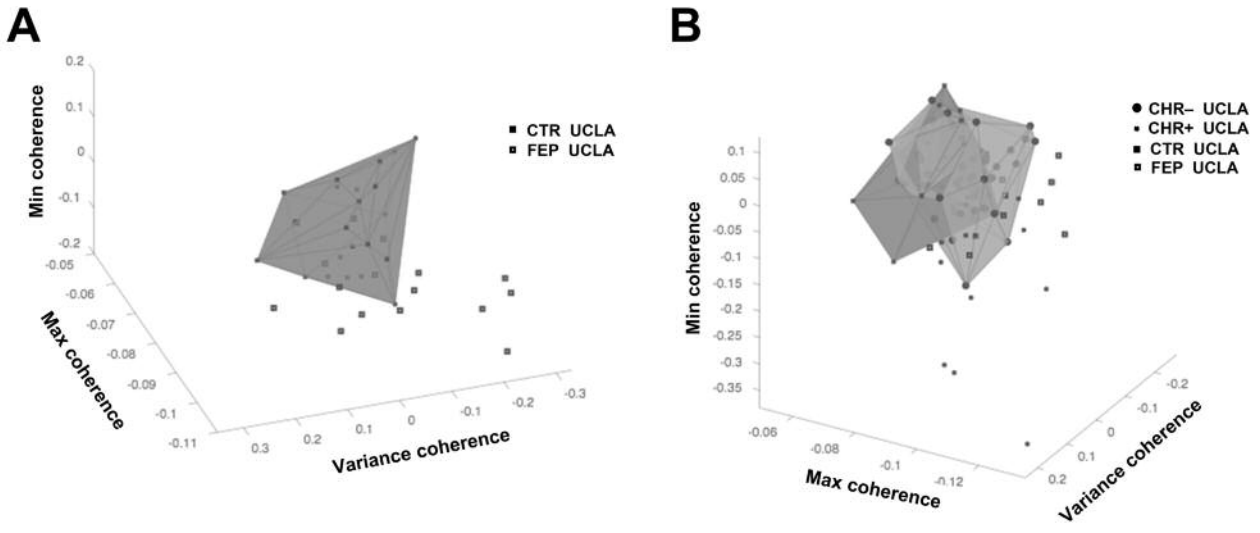

Рис. 4. Проекция трех ведущих факторов для пациентов с первым психотическим эпизодом (FEP) из выборки UCLA и представителей контрольной группы здоровых лиц (CTR). A. Выпуклая оболочка здоровых лиц (CTR) с 11 из 16 пациентов с первым эпизодом за пределами оболочки. B. Перекрытие выпуклых оболочек пациентов с первым психотическим эпизодом и представителей контрольной группы, а также конвертеров CHR+ и не-конвертеров CHR–

Трехфакторное представление разложения векторов на единичные значения продемонстрировало исключение 11 из 16 пациентов с недавно возникшим психозом из выпуклой оболочки, определяемой точками данных здоровых добровольцев, и указало на истинное положительное значение, равное 0,69 (рис. 4A). Было установлено пространственное перекрытие между выпуклыми оболочками, которые содержали данные здоровых лиц и представителей CHR– (рис. 4B).

ОБСУЖДЕНИЕ

Используя автоматизированные методы обработки естественного языка с машинным обучением для анализа речи в когорте CHR, мы создали классификатор, включающий показатели сниженной семантической связности, большей вариабельности этой связности и уменьшения частоты употребления притяжательных местоимений, которые с высокой степенью точности прогнозировали последующее наступление психоза.

Данный классификатор характеризовался внутрипротокольной точностью 83% при использовании в отношении набора данных для обучения, и точностью по перекрестному протоколу – 79% при применении к речевым образцам из второй независимой когорты CHR (тестовый набор данных)[15], демонстрируя значительную сохранность точности прогнозирования, несмотря на различные методы получения образцов речи[8,15]. Кроме того, этот классификатор позволял дифференцировать речь пациентов с недавно начав-

шимся психозом и речь здоровых людей с точностью 72%, что свидетельствует о том, что его дискриминационная сила была относительно надежна на всех стадиях болезни, как было установлено для клинических оценок расстройства мышления[1,6]. Более того, была обнаружена сильная корреляция между прогностически значимыми показателями, полученными с помощью автоматической обработки и кодирования вручную, что доказывает конкурентную валидность подхода.

Было давно замечено, что для речи больных шизофренией характерно нарушение смысловой связности: Крепелин описывал феномен *Sprachverwirrtheit* (например, спутанная речь)[25], Блейлер выделял понятие «разрыхление ассоциаций» в качестве базовой характерной черты шизофрении[26]. Позже Андреасен предложила рассматривать снижение семантической связности как позитивное нарушение мышления[27]. Hoffman использовал ручной дискурс-анализ применительно к транскрибированной речи у больных шизофренией и обнаружил снижение семантической связности[28], что позже было повторено в исследовании с использованием компьютеризированного дискурс-анализа[29].

Только за последнее десятилетие лингвистический анализ естественного языка, в частности латентно-семантический анализ, стали применять для исследования речевой продукции пациентов с шизофренией; он позволил обнаружить снижение семантической связности, которое коррелировало с клиническими оценками, показателями функциональных нарушений и активацией нейронных сетей, отвечающих за язык, в ответ на выполнение заданий[11,12]. На сегодняшний день в двух исследованиях CHR латентно-семантический анализ с применением машинного обучения выявил снижение семантической связности, позволяющее прогнозировать последующее начало психоза.

Нарушение синтаксиса при шизофрении также хорошо задокументировано. Ошибки использования ссылок с местоимениями в речи больных шизофренией были описаны Хоффманом[30] три десятилетия назад, что с тех пор было неоднократно подтверждено в других исследованиях, использовавших методики классификации/подсчета слов[29,31]. В настоящем исследовании, используя пометки (тэги) частей речи, мы выявили снижение использования притяжательных местоимений как прогностический фактор манифестации психоза, который составил большую часть веса четвертого фактора в классификаторе. Полученные данные согласуются с результатами предыдущего лингвистического анализа с применением кодирования вручную, который был выполнен на данных этой же когорты и позволил выявить снижение референциальной связности речи в роли предиктора начала психоза[8], в частности, такое использование местоимений и компаративов («это(т)» или «то(т)»), которое затрудняет понимание того, о ком конкретно или о чем именно пациент говорит.

В речи пациентов с шизофренией часто встречается уменьшение синтаксической сложности[27,32], что обычно выражается в использовании более коротких предложений, и становится наиболее очевидным, когда исследуются расширенные нарративы, записанные без ограничения продолжительности времени повествования[12,30,31,33]. В нашем предыдущем небольшом исследовании естественного языка[15] мы обнаружили две меры синтаксической сложности – более короткие предложения и уменьшение частоты употребления определительных местоимений, использующихся для введения придаточных предложений, – эти показатели не только выступают в качестве предикторов психоза, но и сильно коррелируют с негативными симптомами. В настоящем исследовании ограничение возможности использования показателя длины предложения для прогнозирования манифестации психоза в выборке данных обучения может быть следствием получения кратких и структурированных ответов (<20 слов на ответ)[12], по сравнению с предыдущими исследованиями (>120 слов/ответ[15], 800 слов/ответ[12] и >10 предложений/ответ[30]).

В обоих наших исследованиях CHR мы получили классификации с презентацией в форме выпуклых оболочек, в которых точки речевых данных CHR– (не заболевших лиц с клинически высоким риском психоза) находились внутри корпуса, а точки CHR+ находились за его пределами. Аналогичная выпуклая оболочка была создана для здоровых лиц с помощью классификатора CHR, причем точки речевых данных лиц, недавно перенесших первый психотический эпизод, находились, в основном, за пределами корпуса. Все вместе эти результаты свидетельствуют о том, что речь препсихотических и психотических больных имеет отклонения от корпуса нормального языка в отношении семантики и синтаксиса.

До сих пор нормальные паттерны языка, выделенные с помощью автоматизированных методов обработки естественного языка, остаются недостаточно изученными, в том числе в контексте развития, поскольку семантическая и синтаксическая сложность возрастает в подростковом возрасте и в раннем взрослом возрасте[34]. Следует отметить, что предположение о том, что процессы, лежащие в основе производства нормальной речевой продукции и понимания, являются относительно однородными, поддерживается работами Hasson, которые продемонстрировали выравнивание временных характеристик активации мозга у здоровых людей (межсубъектная связность) как во время слухового восприятия речевой информации, так и во время актов говорения[35].

Обнаружение нами сильных корреляций между автоматизированными и обработанными вручную лингвистическими переменными доказывает конкурентную валидность подхода к обработке естественного языка. Автоматизированные методы обработки естественного языка гораздо быстрее и дешевле, чем ручные лингвистические подходы, и могут быть более легко адаптированы для исследований и, в конечном итоге, для клинических задач.

Помимо методов семантического анализа речи и выделения (тэгирования) частей речи, язык и речь также могут быть оценены с помощью метода составления речевых графов[36], исследования просодии, прагматики, метафоричности[37], а также с помощью анализа дискурса или изучения разговоров с собеседниками. Автоматизированный анализ естественного языка также использовался для исследования характеристик других поведенческих нарушений, включая опьянение наркотическими веществами[38] и болезнь Паркинсона[39], так что данная технология имеет шансы быть использованной в медицине более широко. Наконец, автоматизированные подходы могут быть применены и к другим проявлениям поведения, например к мимике и эмоциям[40]. В целом, автоматизированный анализ речи – это мощная, но при этом недорогая технология, которая может быть использована в психиатрии для диагностики, прогнозирования и оценки терапевтического ответа.

Основные ограничения в настоящем исследовании включают размер выборки и оставшиеся пробелы в наших знаниях относительно того, что является нормой на стадиях развития человека и как нормальная и девиантная речь может быть сопоставлена с обеспечивающими их нейронными сетями. Кроме того, в двух когортах использовались различные методы получения речевого материала, так что связность на уровне предложений не могла быть оценена для первой базы данных обучения из-за краткости ответов участников, что потребовало использование методов «K-уровня» для оценки семантической связности и выравнивающего преобразования данных для проверки перекрестного протокола. В текущих исследованиях мы используем варианты открытых интервью для получения

свободной естественной речи для анализа, чтобы мы могли измерять семантическую связность на уровне предложений и лучше фиксировать показатели синтаксической сложности.

В целом, мы показываем полезность и валидность автоматизированных методов обработки естественного языка для оценки трудноуловимых нарушений ссемантики и синтаксиса на разных стадиях психотического расстройства. Данная технология может повысить качество прогнозирования исхода психоза среди подростков и молодых людей, находящихся в группе высокого риска, и может иметь более значимые последствия для медицинских исследований и практики в целом.

ВЫРАЖЕНИЕ БЛАГОДАРНОСТИ

Данное исследование было поддержано Национальным институтом психического здоровья США (R01 MH 107558; R03 MH 108933 02), Ведомством психического здоровья штата Нью-Йорк, Премией для молодых ученых NARSAD/BBRF и Председателем фонда семьи Миллер C.E. Bearden. Данные источники финансирования не играли никакой роли в проектировании и проведении исследования; сборе, администрировании, анализе и интерпретации данных; подготовке, рассмотрении или одобрении рукописи; решении представить рукопись для публикации.

Библиография

1. Roche E, Creed L, MacMahon D et al. The epidemiology and associated phenomenology of formal thought disorder: a systematic review. Schizophr Bull 2015;41:951-62.
2. Andreasen NC, Grove WM. Thought, language, and communication in schizophrenia: diagnosis and prognosis. Schizophr Bull 1986;12:348-59.
3. Gooding DC, Ott SL, Roberts SA et al. Thought disorder in mid-childhood as a predictor of adulthood diagnostic outcome: findings from the New York High-Risk Project. Psychol Med 2013;43:1003-12.
4. Nelson B, Yuen HP, Wood SJ et al. Long-term follow-up of a group at ultra high risk («prodromal») for psychosis: the PACE 400 study. JAMA Psychiatry 2013;70:793-802.
5. Addington J, Liu L, Buchy L et al. North American Prodrome Longitudinal Study (NAPLS 2): the prodromal symptoms. J Nerv Ment Dis 2015;203:328-35.
6. De Vylder JE, Muchomba FM, Gill KE et al. Symptom trajectories and psychosis onset in a clinical high-risk cohort: the relevance of subthreshold thought disorder. Schizophr Res 2014;159:278-83.
7. Cornblatt BA, Carrion RE, Auther A et al. Psychosis prevention: a modified clinical high risk perspective from the Recognition and Prevention (RAP) program. Am J Psychiatry 2015;172:986-94.
8. Bearden CE, Wu KN, Caplan R et al. Thought disorder and communication deviance as predictors of outcome in youth at clinical high risk for psychosis. J Am Acad Child Adolesc Psychiatry 2011;50:669-80.
9. Landauer TK, Dumais ST. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol Rev 1997;104:211-40.
10. Landauer TK, Foltz PW, Laham D. An introduction to latent semantic analysis. Discourse Process 1998;25:259-84.
11. Elvevag B, Foltz PW, Weinberger DR et al. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. Schizophr Res 2007;93:304-16.
12. Elvevag B, Foltz PW, Rosenstein M et al. An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. J Neurolinguistics 2010;23:270-84.
13. Santorini B. Part-of-speech tagging guidelines for the Penn Treebank Project. 3rd revision. Philadelphia: Department of Computer and Information Science, University of Pennsylvania, 1990.
14. Bird S. Natural language processing and linguistic fieldwork. Comput Linguist 2009;35:469-74.
15. Bedi G, Carrillo F, Cecchi GA et al. Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ Schizophr 2015;1:15030.
16. Miller TJ, McGlashan TH, Rosen JL et al. Prodromal assessment with the structured interview for prodromal syndromes and the scale of prodromal symptoms: predictive validity, interrater reliability, and training to reliability. Schizophr Bull 2003;29:703-15.
17. Caplan R, Guthrie D, Fish B et al. The Kiddie Formal Thought Disorder Rating Scale: clinical assessment, reliability, and validity. J Am Acad Child Adolesc Psychiatry 1989;28:408-16.
18. Ben-David S, Birnbaum ML, Eilenberg ME et al. The subjective experience of youths at clinically high risk of psychosis: a qualitative study. Psychiatr Serv 2014;65:1499-50.
19. Mandera P, Keuleers E, Brysbaert M. How useful are corpus-based methods for extrapolating psycholinguistic variables? Q J Exp Psychol 2015;68:1623-42.
20. Shonemann P. A generalized solution of the orthogonal procrustes problem. Psychometrika 1966;31:1-10.
21. Haxby JV, Guntupalli JS, Connolly AC et al. A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron 2011;72:404-16.
22. Ashburner J, Firston K. Rigid body registration. In: Penny W, Friston K, Ashburner J et al (eds). Statistical parametric mapping: the analysis of functional brain images. Cambridge: Academic Press, 2007:49-62.
23. Mollica RF, Milic M. Social class and psychiatric practice: a revision of the Hollingshead and Redlich model. Am J Psychiatry 1986;143:12-7.
24. Jorge-Botana G, Olmos R, Luzon JM. Word maturity indices with latent semantic analysis: why, when, and where is Procrustes rotation applied? Wiley Interdiscip Rev Cogn Sci (in press).
25. Kraepelin E. Psychiatrie. Ein Lehrbuch fur Studierende und Arzte. Leipzig: Barth, 1899.
26. Bleuler E. Dementia Praecox oder Gruppe der Schizophrenien. Leipzig: Deuticke, 1911.
27. Andreasen NC. Thought, language, and communication disorders. I. Clinical assessment, definition of terms, and evaluation of their reliability. Arch Gen Psychiatry 1979;36:1315-21.
28. Hoffman RE, Stopek S, Andreasen NC. A comparative study of manic vs schizophrenic speech disorganization. Arch Gen Psychiatry 1986;43:831-8.
29. Noel-Jorand MC, Reinert M, Giudicelli S et al. A new approach to discourse analysis in psychiatry, applied to a schizophrenic patient's speech. Schizophr Res 1997;25:183-98.
30. Hoffman RE, Hogben GL, Smith H et al. Message disruptions during syntactic processing in schizophrenia. J Commun Disord 1985;18:183-202.
31. Buck B, Penn DL. Lexical characteristics of emotional narratives in schizophrenia: relationships with symptoms, functioning, and social cognition. J Nerv Ment Dis 2015;203:702-8.
32. Kuperberg GR. Language in schizophrenia Part 2: What can psycholinguistics bring to the study of schizophrenia... and vice versa? Lang Linguist Compass 2010;4:590-604.
33. Andreasen NC. Thought, language, and communication disorders. II. Diagnostic significance. Arch Gen Psychiatry 1979;36:1325-30.
34. Nippold MA, Ward-Lonergan JM, Fanning JL. Persuasive writing in children, adolescents, and adults: a study of syntactic, semantic, and pragmatic development. Lang Speech Hear Serv Sch 2005;36:125-38.
35. Silbert LJ, Honey CJ, Simony E et al. Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. Proc Natl Acad Sci USA 2014;111:E4687-96.
36. Mota NB, Vasconcelos NA, Lemos N et al. Speech graphs provide a quantitative measure of thought disorder in psychosis. PLoS One 2012;7:e34928.
37. Gutierrez ED, Shuotva E, Marghetis T et al. Literal and metaphorical senses in compositional distributional semantic models. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, 2016:183-93.
38. Bedi G, Cecchi GA, Slezak DF et al. A window into the intoxicated mind? Speech as an index of psychoactive drug effects. Neuropsychopharmacology 2014;39:2340-8.
39. Garcia AM, Carrillo F, Orozco-Arroyave JR et al. How language flows when movements don't: an automated analysis of spontaneous discourse in Parkinson's disease. Brain Lang 2016;162:19-28.
40. Baker JT, Pennant L, Baltrusaitis T et al. Toward expert systems in mental health assessment: a computational approach to the face and voice in dyadic patient-doctor interactions. iproc 2016;2:e44.

DOI: 10.1002/wps.20491