Psychological Medicine

cambridge.org/psm

Original Article

Cite this article: Barroilhet SA, Pellegrini AM, McCoy TH, Perlis RH (2019). Characterizing DSM-5 and ICD-11 personality disorder features in psychiatric inpatients at scale using electronic health records. *Psychological Medicine* 1–9. https://doi.org/10.1017/S0033291719002320

Received: 26 April 2019
Revised: 9 August 2019
Accepted: 14 August 2019

Key words:
Electronic health record; length of stay; machine learning; natural language processing; personality disorder

Author for correspondence:
Roy H. Perlis,
E-mail: rperlis@partners.org

CAMBRIDGE
UNIVERSITY PRESS

Characterizing DSM-5 and ICD-11 personality disorder features in psychiatric inpatients at scale using electronic health records

Sergio A. Barroilhet[1,2,3], Amelia M. Pellegrini[1], Thomas H. McCoy[1] and Roy H. Perlis[1]

[1]Center for Quantitative Health, Division of Clinical Research and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA; [2]Department of Psychiatry, Tufts University School of Medicine, Boston, MA, USA and [3]University Psychiatric Clinic, University of Chile Clinical Hospital, Santiago, Chile

Abstract

Background. Investigation of personality traits and pathology in large, generalizable clinical cohorts has been hindered by inconsistent assessment and failure to consider a range of personality disorders (PDs) simultaneously.

Methods. We applied natural language processing (NLP) of electronic health record notes to characterize a psychiatric inpatient cohort. A set of terms reflecting personality trait domains were derived, expanded, and then refined based on expert consensus. Latent Dirichlet allocation was used to score notes to estimate the extent to which any given note reflected PD topics. Regression models were used to examine the relationship of these estimates with sociodemographic features and length of stay.

Results. Among 3623 patients with 4702 admissions, being male, non-white, having a low burden of medical comorbidity, being admitted through the emergency department, and having public insurance were independently associated with greater levels of disinhibition, detachment, and psychoticism. Being female, white, and having private insurance were independently associated with greater levels of negative affectivity. The presence of disinhibition, psychoticism, and negative affectivity were each significantly associated with a longer stay, while detachment was associated with a shorter stay.

Conclusions. Personality features can be systematically and scalably measured using NLP in the inpatient setting, and some of these features associate with length of stay. Developing treatment strategies for patients scoring high in certain personality dimensions may facilitate more efficient, targeted interventions, and may help reduce the impact of personality features on mental health service utilization.

Introduction

Personality disorder (PD) diagnoses have an important public health impact as they predict increased utilization of medical and mental health care services (Twomey *et al.*, 2015; Tyrer *et al.*, 2015; Huprich, 2018). Studies using structured diagnostic interviews have identified a PD diagnosis in 40–82% of psychiatric outpatient populations (Zimmerman *et al.*, 2005; Newton-Howes *et al.*, 2010; Beckwith *et al.*, 2014) and in 64–74% of psychiatric inpatient populations (Grilo *et al.*, 1998; Keown *et al.*, 2005; Stevenson *et al.*, 2011), further increasing utilization in these settings (Twomey *et al.*, 2015).

The variability in these prevalence estimates suggests the challenge of studying PDs in real-world settings. Despite high levels of usage of health care resources, and high rates of polypharmacy and hospital admissions (Quirk *et al.*, 2016) and the economic burden associated (Soeteman *et al.*, 2008), evaluating personality dimensions is still not a part of routine assessment in psychiatric inpatient units (Fok *et al.*, 2014; Jacobs *et al.*, 2015). Likewise, in administrative data sets, PDs may not be coded consistently, or may be treated as a single undifferentiated category (Jiménez *et al.*, 2004; McLay *et al.*, 2005; Compton *et al.*, 2006; Jacobs *et al.*, 2015; Newman *et al.*, 2018). On the other hand, the current categorical diagnosis for PDs has been questioned as not scientifically valid, while PD clinical features are being increasingly understood as dimensional phenotypes (Bjelland *et al.*, 2009; Haslam *et al.*, 2012; Skodol, 2012; Tyrer *et al.*, 2015). Accordingly, the DSM-5 and ICD-11 have both moved toward dimensional models of PD (Bach *et al.*, 2018a,b) and remain to be studied. Novel approaches to explore personality dimensions in psychiatric cohorts are needed (Quirk *et al.*, 2016).

To address this gap, we applied natural language processing (NLP) of electronic health records (EHRs) to characterize a large inpatient psychiatric cohort (Manning and Schiitze, 1999). We hypothesized that EHR notes would capture relevant clinical descriptions as unstructured data, quantifiable by validated algorithmic tools that have been previously used for medical (Yu *et al.*, 2014; Yim *et al.*, 2016) and mental health research (Althoff *et al.*, 2016; Can *et al.*, 2016; McCoy *et al.*, 2016; Birnie *et al.*, 2018; McCoy *et al.*, 2018; Afshar *et al.*, 2019). In particular, we examined the relationship between these dimensions and sociodemographic and clinical features, as a means of more comprehensively characterizing personality psychopathology in a real-world setting.

Methods

Subjects

Sociodemographic and clinical data were extracted from the health records of patients in the adult psychiatry inpatient unit at Massachusetts General Hospital between 2010 and 2016. Sociodemographic data included age, sex, race, and type of insurance, as well as relevant clinical factors such as admission route (i.e. either via the emergency room or not), length of stay, and Charlson Comorbidity Index. Admission and discharge documentation were extracted for estimation of personality trait domains by NLP. These EHR data were managed as an i2b2 datamart (Murphy *et al.*, 2010).

The Partners HealthCare Human Research Committee approved the study protocol, waiving the requirement for informed consent as detailed by 45 CFR 46.116 as no participant contact was required in this study based on secondary use of data arising from routine clinical care.

Generation of personality phenotypes

Building on our prior work in transdiagnostic psychiatric phenotypes, we developed personality-specific transdiagnostic phenotypes based on NLP (McCoy *et al.*, 2018). This process seeds an NLP model using expert-defined, or curated, terms. As with our prior work, we consulted relevant texts to guide phenotypic seed term generation; in this case, the DSM-5 and ICD-11. The DSM-5 (section III) (American Psychiatric Association, 2013) and ICD-11 (Tyrer *et al.*, 2015; Bach and First, 2018) assess PDs based on determining levels of functioning/impairment and stylistic traits organized in personality dimensions. In the DSM-5, these dimensions are Negative Affectivity, Detachment, Antagonism, Disinhibition, and Psychoticism. The ICD-11 includes the same dimensions, except Psychoticism, and adds Anankastia (or Compulsivity) as a new dimension. Definitions of overlapping dimensions are similar between the DSM-5 and ICD-11 (Bach *et al.*, 2018b). These extracted trait domain definitions, according to Skodol (2018) and Tyrer *et al.* (2015), are shown in Table 1 along with the examples of personality features that comprise these dimensions. These DSM-5 and ICD-11 derived terms were then expanded using the Personality Inventory for DSM-5 items (Krueger *et al.*, 2012), other personality trait studies (Ashton *et al.*, 2004, 2012; Bach *et al.*, 2018*a*, 2018*b*), and a thesaurus (Dictionary.com, LLC, 2019). From the generated synonym list, a clinically refined set of NLP seed terms was selected based on expert consensus (S.A.B., R.H.P.; Table 1).

As these pre-selected term lists are unlikely to capture the full diversity of clinical vocabulary, we applied a previously reported method for expanding clinical vocabularies (McCoy *et al.*, 2018). In this method, Latent Dirichlet allocation (LDA) is used to fit a probabilistic topic model to all documents. The use of topic loadings as LDA-determined phenotypes has been used for computational phenotyping and is discussed in our prior research (McCoy *et al.*, 2017, 2018). Briefly, with an LDA-based topic model, documents are probability distributions over topics, and each topic is a probability distribution over the full vocabulary (Blei *et al.*, 2003; Blei, 2012). The posterior distributions of the term-topic distributions are inspected to identify the topic under which the cumulative probability of the expert-selected personality token within each list is greatest. This total cumulative probability of the seed word list is used to identify the relevant topic. Thereafter, that topic's topic-document weights are used as the phenotype for the relevant domain. In essence, this approach asks which LDA topics capture the greatest number of curated tokens for a given PD, and then uses the 'best' topic to represent that disorder. The tokens (terms) incorporated in topics corresponding to each concept are listed in Table 1, and the entire process is outlined in Fig. 1. For the topic modeling, we used the R interface to a Gibbs sampler implementation of LDA (*topicmodels* v0.2), one of many widely used open source implementations of LDA licensed under free software licenses (McCallum, 2002; Řehůřek and Sojka, 2010; Grün and Hornik, 2018).

Study design and analysis

We used robust clustering to account for individuals with multiple admissions. Linear regression modeling adjusting for sex, age, race, insurance type, Charlson Comorbidity Index, and route of admission was used to analyze personality domain loadings in different sociodemographic profiles. Linear regression adjusting for these sociodemographic variables, as well as for other personality trait domains, was used to explore the association between personality trait domains and hospital length of stay. Analyses utilized Stata/SE 13.1 (Statacorp, College Station, TX, USA).

Results

Characteristics of the full set of 4702 admissions for 3623 individuals are displayed in Table 2. Individual personality trait domains differed in their association with sociodemographic features (Table 3). Being male, non-white, having a low burden of medical comorbidity, being admitted through the emergency room, and having public insurance were independently associated with higher levels of disinhibition, detachment, and psychoticism. On the other hand, being female, white, and using private insurance were independently associated with increased levels of negative affectivity. Age was also associated with personality features: on average, patients with increased levels of disinhibition and psychoticism were younger, while patients with more negative affectivity were older.

We next examined the association between personality trait domains extracted from clinical notes and length of inpatient stay. As shown in Table 4, the presence of disinhibition, psychoticism, and negative affectivity was significantly associated with a longer length of stay. In contrast, detachment was associated with a shorter length of stay. A 10% increase in the disinhibition domain score was associated with a ~2.7-day increase in length of stay. Similarly, a 10% increase in the psychoticism and negative affectivity domain scores was associated with an increase in length of stay of ~0.8 and ~0.7 days, respectively. On the other hand, having a 10% increase in

Table 1. Personality trait domains in DSM-5 and ICD-11

Personality trait domain	Diagnostic system	Definition	Main personality features	Personality traits used as topics
Negative affectivity	DSM-5 and ICD-11	Frequent and intense experiences of high levels of a wide range of negative emotions (e.g. anxiety, depression, guilt/shame, worry, anger, etc.), and their behavioral (e.g. self-harm) and interpersonal (e.g. dependency) manifestations	Depressed, pessimistic, remorseful, anxious, worried, submissive, dependent	Depressive, depressing, depressed, depression, overwhelmed, disheartened, dispirited, discouraging, gloomy, glum, downcast, cheerless, dim, hopeless, hopelessness, melancholic, melancholy, dismal, despaired, despair, discouraged, discouraging, despondent, sadness, sad, sorry, remorse, remorseful, unhappy, disconsolate, miserable, oppressed, sunk, sunken, blue, blues, dejected, misery, sorrowful, unlucky, sorrow, spiritless, desolate, grim, sobbing, ripped, cry, crying, weeping, nostalgic, regretful, longing, yearning, homesick, negative, fatalistic, dissatisfaction, dissatisfied, demoralized, disappointing, disappointed, frustrated, frustration, pessimism, pessimistic, resignation, resigned, resign, guilt, pitiful, guilty, anxious, troubled, worrying, vulnerable, moody, emotional, insecure, frightened, afraid, apprehensive, careful, concerned, distress, distressed, fearful, fidgety, restless, scare, scared, uneasy, jumpy, nervy, nervous, shivery, unquiet, worried, worry, jittery, uptight, clutched, disturbing, disturbed, dreading, tensed, tense, sensitive, anguish, shy, spooked, neurotic, edgy, unrestful, unsettled, inadequate, loose, lost, looser, fruitless, useless, weary, blunt, dull, jealous, instability, unstable, oversensitive, hypersensitive, panic, panicky, indecision, indecisive, sentimental, fragile, touchy, whining, complaining, suggestible, hesitating, victimization, victim, influenceable, irresolute, weakness, weak, doubtful, coward, timorous, timid, bashful, submissiveness, submissive
Antagonism/ dissociality	DSM-5 and ICD-11	Behaviors that put the individual at odds with other people, including an exaggerated sense of self-importance and a concomitant expectation of special treatment, as well as a callous antipathy toward others, encompassing both unawareness of others' needs and feelings, and a readiness to use others in the service of self-enhancement	Manipulative, deceitful, grandiose, callous, hostile, violent	Mean, pretentious, impolite, treacherous, unfairness, unfair, hypocritical, grasping, boastful, corrupt, false, lie, liar, lying, dishonest, smug, greedy, haughty, ostent, ostentatious, snob, snobbish, conceited, antagonistic, selfish, rude, coldness, cold, suspiciousness, suspicious, vengeful, revengeful, retaliatory, manipulative, arrogant, callous, grandiosity, grandiose, hostility, hostile, deceitful, exploitative, egocentric, calculating, devious, conniving, unscrupulous, disingenuous, rebuffing, rejective, noncompliant, hesitant, unwilling, disobedience, disobedient, insubordinate, subversive, rebellious, rebel, rebellious, disruptive, defiant, aggression, aggressive, violence, violent, ruthless agitation, agitated, critic, critical. Antagonistic, selfish, fierce, mutinous, explosion, explosive, bossy, authoritarian, hurtful, brusque, choleric, hard, irritable, irritability, rough, provoking, tyrant, tyrannical, egoistic, egotistical, egotist, pitiless, litigious, bellicose, overbearing, oppressive, intolerance, intolerant, angry, irascibility, testiness, testy, tetchy, irascible, quarrelsome, surly, fundamentalist, polemical, dogmatic, extremist, heartless, harsh, vehement, disputatious, roistering, excitable, pompous, pretending, stingy, deceiving, insincere, miserly, avaricious, disloyal, untruthful, venal, gossip, gossipy, malicious, betraying, boasting, flattering, mercenary, sly, vindictive, envious

(*Continued*)

Table 1. (*Continued.*)

Personality trait domain	Diagnostic system	Definition	Main personality features	Personality traits used as topics
Disinhibition	DSM-5 and ICD-11	Orientation toward immediate gratification, leading to impulsive behavior driven by current thoughts, feelings, and external stimuli, without regard for past learning or consideration of future consequences	Irresponsible, impulsive, distractible, reckless, thoughtless	Distractible, impetuous, hasty, impatient, reckless, thoughtless, risky, careless, unconscious, foolhardy, daredevil, dare, daring, brash, overbold, impulsive, capricious, immature, feckless, fickle, flighty, harebrained, incautious, lax, scatterbrained, uncareful, unpredictable, sloppy, inefficient, negligent, neglected, laziness, lazy, irresponsible, aimless, unreliable, indolent, licentious, frivolous, untidy, disorderly, languid, idle, inconstant, imprecise, imprudent, irrational, rambling, undisciplined, unreflecting, dissolute, bungling, inaccurate, unfaithful, inattentive, silly, childish, inconsiderate, unwise, inconsequent, unpersevering
Detachment	DSM-5 and ICD-11	Avoidance of socioemotional experience, including both withdrawal from interpersonal interactions (ranging from casual, daily interactions to friendships to intimate relationships) and restricted affective experience and expression, particularly limited hedonic capacity	Detached, unmotivated, insensitive, reserved, avoidant, isolated, aloof	Aloof, retiring, insensitive, withdrawn, avoidant, anhedonic, detached, isolated, reserved, distant, unsociable, remote, unapproachable, uncommunicative, introverted, restrained, retired, retreated, shrinking, disinterested, incurious, indifferent, non-gregarious, offish, recluse, reclusive, silent, solitary, standoffish, taciturn, uncompanionable, unconcerned, undemonstrative, unforthcoming, uninterested, apathetic, diffident, impervious, nonchalant, uncaring, uninvolved, unresponsive, unsympathetic, dispassionate, heedless, listless, nonpartisan, passionless, phlegmatic, scornful, unaroused, unemotional, unmoved, unprejudiced, unsocial, laconic, reticent, curt, unempathetic, dumb, speechless, unexpressive, apart, secluded, away, sequestered, candid, impersonal, lackadaisical
Psychoticism	Only DSM-5	Exhibiting a wide range of culturally incongruent, odd, eccentric, or unusual behaviors and cognitions, including both thought process (e.g. perception, dissociation) and content (e.g. beliefs)	Eccentric, abnormal, odd, bizarre, strange, weird	Eccentric, unconventional, zany, madcap, peculiar, odd, strange, bizarre, weird, crazy, extravagant, freak, grotesque, unusual, singular, wild, ludicrous, ridiculous, outlandish, flamboyant, awkward, awry, dowdy, erratic, idiosyncratic, kooky, offbeat, quirky, whimsical, aberrant, nutty, bent, oddball, anomalous, cockeyed, freakish, funky, quaint, quizzical, uncommon, unreasonable
Anankastia	Only ICD-11	Narrow focus on the control and regulation of one's own and others' behavior to ensure that things conform to the individual's particularistic ideal. Traits in this domain include concern with following rules and meeting obligations	Rigid, perfectionist, perseverative, obsessive, stubborn, controlling	Dutiful, disciplined, punctual, scrupulous, neat, persevering, organized, ambitious, meticulous, precise, orderly, industrious, thorough, tidy, studious, perfection, perfectionistic, methodical, rigorous, faultless, inartistic, traditional, conventional, businesslike, systematic, demanding, fussy, exacting, rigid, overproductive, inflexible, moralistic, insistent, flawless, perfect, detailed, intransigent, stern, stringent, stubborn, headstrong, obstinate, fixed, unyielding, bullheaded, changeless, obdurate, strict, stiffness, stiff, choosy, finicky, squeamish, exact, painstaking, obsessive, obsessiveness, punctilious, querulous, stickling

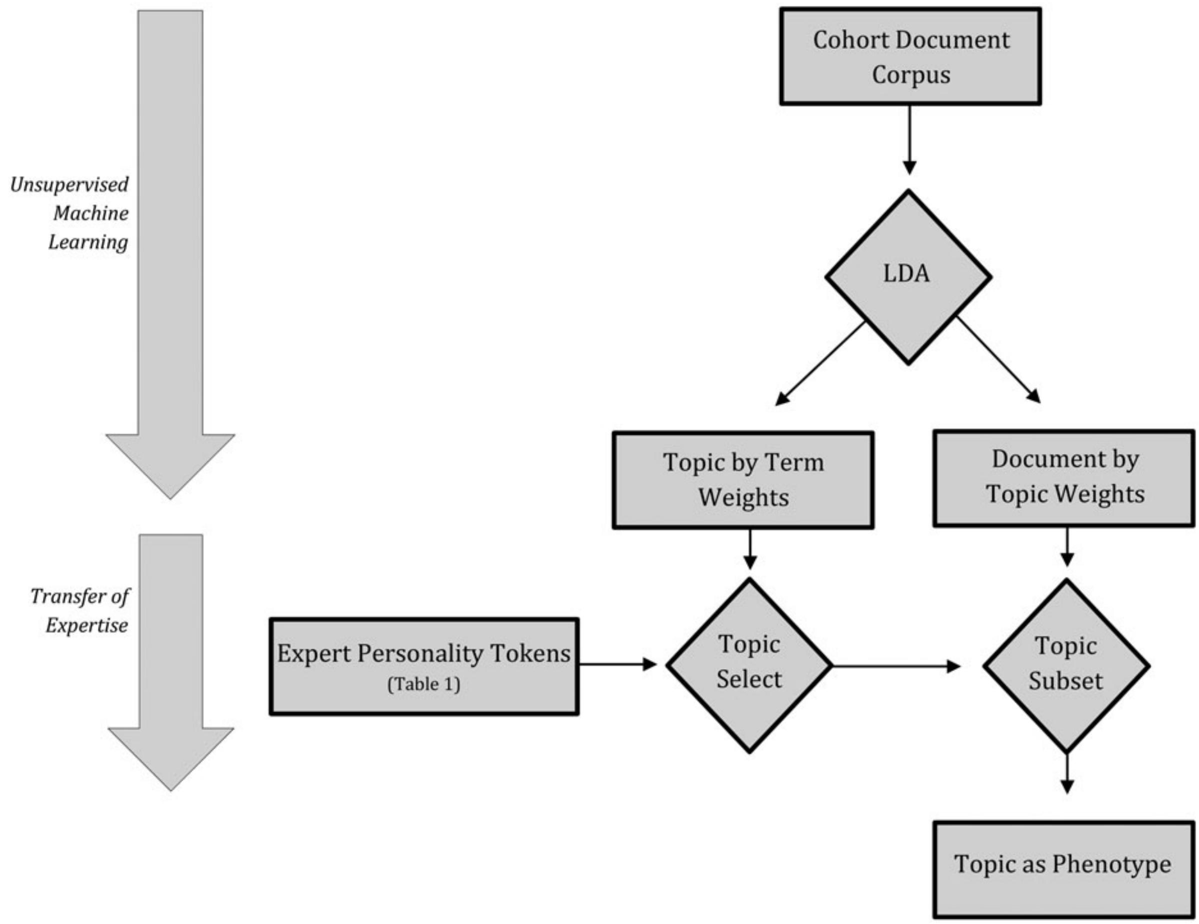

Fig. 1. Diagram of personality phenotype generation through the transfer of human expert language model into model learned through unsupervised machine learning. A probabilistic topic model is learned from patients' clinical documentation through LDA. The learned topics are then matched to the personality symptom domains by linking the learned topic under which expert-identified tokens are most common. Thereafter, the linked topics are used as the phenotype for the linked personality domain.

detachment features was associated with a decreased length of stay by nearly 0.3 days.

Discussion

As anticipated based on studies using traditional personality measures, we observed an association between sociodemographic features and individual personality trait domains (Lynn and Martin, 1997; Kjelsås and Augestad, 2004). Demographic profiles are useful to predict certain behaviors (Krismayer *et al*., 2019), but their relationship with dimensional traits is less studied (Al-Halabí *et al*., 2010).

In particular, we found that greater scores in disinhibition, negative affectivity, and psychoticism were associated with a significantly longer length of stay, while a greater score in detachment was associated with a decreased length of stay. One way to interpret the effect sizes we observed is to compare our results to the US national average length of stay in inpatient psychiatric units, which is 6.6 days (Heslin *et al*., 2015). According to our results, an increase of 10% in the disinhibition dimension score may increase inpatient length of stay by 40% when compared to the national average. Likewise, patients scoring 10% higher in either psychoticism or negative affectivity may have an increased length of stay by an extra 12% when compared to the national average. Conversely, an increase of 10% in the detachment dimension score may decrease length of stay by 6% when compared to the national average. Given these results, personality may be a relevant factor to consider in terms of length of stay in the psychiatric inpatient setting.

While there is no doubt that PDs in general are associated with an increase in mental health services use in the outpatient setting (Twomey *et al*., 2015; Tyrer *et al*., 2015), this relationship has been less clear in terms of psychiatric inpatient services use. In contrast to our results, several epidemiological studies (Jacobs *et al*., 2015; Piccinelli *et al*., 2016; Pauselli *et al*., 2017; Newman *et al*., 2018) and service use studies (Jiménez *et al*., 2004; McLay *et al*., 2005; Compton *et al*., 2006; Leontieva and Gregory, 2013; Habermeyer *et al*., 2018) have shown that PDs do not necessarily increase, and may even shorten, length of stay. Consequently, personality may have been overlooked as an addressable factor in efforts to optimize services use. Only a few studies have found that personality was associated with an increased length of stay (Tyrer and Simmonds, 2003; Fok *et al*., 2014). However, neither of these studies explored which personality traits or diagnosis was associated with this outcome.

Table 2. Cohort characteristics at admission

Variables	$N = 4702$
Age at discharge [years, mean (SD)]	44.97 (16.64)
Length of stay [mean (SD)]	9.95 (3.96)
Log Charlson Comorbidity Index [mean (SD)]	3.14 (3.96)
Sex [male, n (%)]	2327 (49.37)
Public insurance [n (%)]	2873 (60.96)
Admission through emergency room [n (%)]	3130 (66.41)
Race/ethnicity [n (%)]	
White	3412 (72.40)
Black	466 (9.89)
Hispanic	395 (8.38)
Asian	176 (3.73)
Other	253 (5.37)
Diagnosis at admission [n (%)]	
Major depressive disorder	1021 (21.73)
Bipolar disorder	605 (12.88)
Other mood disorders	510 (10.85)
Schizophrenia	432 (9.19)
Other psychosis	389 (8.28)
Substance use disorders	146 (3.11)
Anxiety and other neurotic disorders	121 (2.58)
Personality disorders	26 (0.55)
Other disorders	1449 (30.84)

The only prior study we identified that similarly investigated the association between different personality types and use of services in the psychiatric inpatient setting is Keown et al. (2005). This study considered a cohort of 193 patients from a community served by a mental health team in the UK, who were assessed using a structured interview, diagnosed according to the ICD-10, and followed over a 4-year period. Keown *et al.* found that among non-psychotic patients, having paranoid, dependent, and emotionally unstable PD was associated with an increased length of stay by 150 days in the 4-year period when a patient had one PD disorder, and up to 321 days for patients who had two or all of these PD disorders. Among psychotic patients, length of stay was associated with having more paranoid and anxious traits. Conversely, in the latter group of psychotic patients, the presence of anankastic traits was associated with a shorter length of stay.

The results of the Keown *et al.* study are in line with those from our study, since there is evidence of a correspondence between unstable personality and disinhibition, between paranoid personality and psychoticism, and between anxiety/dependence and negative affectivity (Skodol, 2018). On the other hand, unlike the Keown *et al.* study, we found that detachment – and not anankastic traits – was associated with a shorter length of stay. Interpersonal distance and restriction in the expression of affect may be associated with diminished expression of need for care, so when behavioral symptoms remit, these patients may be more likely to be discharged. However, the anankastia domain

Table 3. Association between sociodemographic features and personality trait domains§

	Negative affectivity	Antagonism	Disinhibition	Detachment	Psychoticism	Anankastia
	β (95% CI)[a]	β (95% CI)[a]	β (95% CI)[a]	β (95% CI)[a]	β (95% CI)[a]	β (95% CI)[a]
Age at discharge	0.0012*** (0.0010 to 0.0013)	0.0001* (8.2e-06 to 0.0001)	−0.0005*** (−0.0006 to −0.0002)	9.89e-07 (−0.0002 to 0.00017)	−0.0001** (−0.0002 to −0.00002)	0.0001 (−0.00001 to 0.0004)
Sex, male	−0.01471*** (−0.0185 to −0.01088)	−0.0018 (−0.0038 to 0.0001)	0.0139*** (0.0085 to 0.0192)	0.0055* (0.0009 to 0.0101)	0.0052*** (0.0029 to 0.0076)	0.0034 (−0.0033 to 0.0101)
Race, white	0.0099*** (0.0062 to 0.0137)	0.0050*** (0.0029 to 0.0070)	−0.0341*** (−0.0413 to −0.02686)	−0.0085** (−0.0143 to −0.0027)	−0.0076*** (−0.0106 to −0.0047)	0.0039 (−0.0036 to 0.0113)
Public insurance	−0.0084*** (−0.0119 to −0.0049)	0.0018 (−0.0038 to 0.0001)	0.01343** (0.0081 to 0.0187)	0.0062** (0.0014 to 0.0111)	−0.0034*** (−0.0059 to −0.0010)	0.0053 (−0.0016 to 0.0123)
Admission through ER	−0.0009 (−0.0047 to 0.0030)	0.0003 (−0.0015 to 0.0020)	0.0085*** (0.0038 to 0.0133)	−0.0065** (−0.0114 to −0.0016)	0.0052*** (0.0032 to 0.0073)	−0.1189*** (−0.1274 to −0.1103)
Charlson Comorbidity Index	0.0003 (−0.0004 to 0.0010)	−0.0003* (−0.0006 to −0.00003)	−0.0022*** (0.0038 to 0.0133)	−0.0009** (−0.0016 to −0.0002)	−0.0010*** (−0.0012 to −0.0007)	−0.0001 (−0.0011 to 0.0010)

CI, confidence interval; ER, emergency room.
[a] β (95% confidence interval) is equal to the variation (and its 95% CI) in days of length of stay, if the named personality domain score increased/decreased by 10%.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Table 4. Regression model of personality trait domains and hospital length of stay ($n = 4687$ admissions)

Variables	Model with personality trait domains	
	β (95% confidence interval)[a]	p-value
Personality domain		
Disinhibition	2.705 (2.182 to 3.228)	<0.001
Psychoticism	0.802 (0.188 to 1.416)	0.010
Negative affectivity	0.723 (0.231 to 1.215)	0.004
Antagonism/Dissociality	0.571 (−0.355 to 1.499)	0.227
Anankastia	−0.116 (−0.325 to 0.093)	0.277
Detachment	−0.290 (−0.566 to −0.015)	0.039
Sociodemographic features		
Age at discharge	0.009 (0.007 to 0.012)	<0.001
Sex, male	−0.088 (−0.144 to −0.032)	0.002
Race, white	0.037 (−0.027 to 0.102)	0.262
Public insurance	−0.001 (−0.053 to 0.054)	0.981
Admission through ER	−0.154 (−0.218 to −0.089)	<0.001
Charlson Comorbidity Index	−0.007 (−0.015 to 0.001)	0.094

ER, emergency room.
[a]β (95% confidence interval) is equal to the variation (and its 95% CI) in days of length of stay, if the named personality trait domain score increased/decreased by 10%.

also shows a correlation with detachment (Skodol, 2018), ranging from 0.46 (Bach *et al.*, 2018b) to 0.79 (Lugo *et al.*, 2019).

Limitations

There are several limitations of our study to be considered. Extracting personality trait domains from EHR notes of psychiatry inpatients is limited by the fact that topics identified by the NLP process may account for state-related symptoms in the context of acute psychiatric syndromes like depression or psychosis, and not for stable personality traits. However, some studies show that trait assessments established during acute episodes (e.g. a major depressive episode) may be valid reflections of personality pathology rather than artifacts of symptomatic state (Morey *et al.*, 2010; Sevilla-Llewellyn-Jones *et al.*, 2017). Another alternative is that personality may itself influence symptom expression, and hence a clinical feature may be an expression of both symptoms and traits (von Gunten *et al.*, 2009; Widiger, 2011). Personality dimensions and common psychiatric disorders also covary (Wright and Simms, 2015) and may be part of spectra, that is, larger constellations of syndromes sharing some common features (Kotov *et al.*, 2017). The approach taken here does not distinguish trait from state effects, but it may still capture relevant clinical features at a given point in time.

Conversely, this study does address several key limitations in the prior evidence base. First, personality diagnosis tends to be overlooked by clinicians; some studies indicate that PD prevalence may be underestimated in psychiatry inpatient settings (Fok *et al.*, 2014; Jacobs *et al.*, 2015), especially in the absence of structured assessments (Zimmerman *et al.*, 2008; Leontieva and Gregory, 2013; Newman *et al.*, 2018). Second, when using only a clinical diagnostic approach, there may be a variation regarding which disorders are more likely to be diagnosed and which may be overlooked. This may be based on factors such as symptom severity, expectation of response to treatment, or familiarity with particular PD diagnoses (Zimmerman and Morgan, 2013; Zimmerman, 2016). Finally, most prior personality studies have used a categorical diagnostic approach for PD diagnosis, which has been criticized for its questionable validity (Haslam *et al.*, 2012; Skodol, 2012; Tyrer *et al.*, 2015).

To address these limitations, we used NLP and machine learning as a novel method to overcome underdiagnosis, selective diagnosis, and lack of characterization of personality in the inpatient setting. In particular, our methodology allows access to extensive clinical information, deliberations, and clinicians' clinical judgment that may not be reflected in coded diagnoses. Likewise, we used a dimensional model to assess personality, in contrast to previous studies that used a categorical approach. This method may account more realistically for specific and clinically significant personality features in the inpatient setting.

Conclusion

In aggregate, our study suggests that personality features can be systematically and scalably measured using NLP in the inpatient setting, and that these features may relevantly contribute to service utilization. Developing treatment strategies for patients scoring high in PD features may facilitate more efficient, targeted interventions, and may help reduce the impact on mental health service utilization.

Author contributions. Drs. Barroilhet and Perlis designed the study, analyzed the results, and drafted the manuscript. Dr. McCoy developed the natural language processing methodology and revised the manuscript. Ms. Pellegrini revised the manuscript. All authors have given final approval of this submission.

Financial support. This work was supported by the National Institute of Mental Health (R.H.P., grant number 1R01MH106577). The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Conflict of interest. Dr. Barroilhet and Ms. Pellegrini report no conflicts of interest. Dr. McCoy receives research funding from the Brain and Behavior Research Foundation, National Institute of Aging, Telefonica Alfa, and The Stanley Center at the Broad Institute. Dr. Perlis holds equity in Psy Therapeutics and Outermost Therapeutics; serves on the scientific advisory boards of Genomind and Takeda; and consults to RID Ventures. Dr. Perlis receives research funding from NIMH, NHLBI, NHGRI, and Telefonica Alfa. Dr. Perlis is an associate editor for *JAMA-Network Open*.

Ethical standards. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. The study protocol has been approved by the Partners HealthCare Human Research Committee (protocol number 2016P002084). The requirement for informed consent was waived as detailed by 45 CFR 46.116 since no participant contact was required in this study based on secondary use of data arising from routine clinical care.

References

Afshar M, Phillips A, Karnik N, Mueller J, To D, Gonzalez R, Price R, Cooper R, Joyce C and Dligach D (2019) Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation. *Journal of the American Medical Informatics Association: JAMIA* **26**, 254–261.

Al-Halabí S, Herrero R, Saiz PA, Garcia-Portilla MP, Corcoran P, Teresa Bascaran M, Errasti JM, Lemos S and Bobes J (2010) Sociodemographic factors associated with personality traits assessed through the TCI. *Personality and Individual Differences* **48**, 809–814.

Althoff T, Clark K and Leskovec J (2016) Large-scale analysis of counseling conversations: an application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics* **4**, 463–476.

American Psychiatric Association (2013) *Diagnostic and Statistical Manual of Mental Disorders (*DSM-5®*)*, 5th Edn. Washington, DC: American Psychiatric Association.

Ashton MC, Lee K, Perugini M, Szarota P, de Vries RE, Di Blas L, Boies K and De Raad B (2004) A six-factor structure of personality-descriptive adjectives: solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology* **86**, 356–366.

Ashton MC, Lee K, de Vries RE, Hendrickse J and Born MPH (2012) The maladaptive personality traits of the personality inventory for *DSM*-5 (PID-5) in relation to the HEXACO personality factors and schizotypy/dissociation. *Journal of Personality Disorders* **26**, 641–659.

Bach B and First MB (2018) Application of the ICD-11 classification of personality disorders. *BMC Psychiatry* **18**, 351.

Bach B, Sellbom M and Simonsen E (2018a) Personality inventory for *DSM*-5 (PID-5) in clinical versus nonclinical individuals: generalizability of psychometric features. *Assessment* **25**, 815–825.

Bach B, Sellbom M, Skjernov M and Simonsen E (2018b) ICD-11 and DSM-5 personality trait domains capture categorical personality disorders: finding a common ground. *Australian & New Zealand Journal of Psychiatry* **52**, 425–434.

Beckwith H, Moran PF and Reilly J (2014) Personality disorder prevalence in psychiatric outpatients: a systematic literature review. *Personality and Mental Health* **8**, 91–101.

Birnie KI, Stewart R and Kolliakou A (2018) Recorded atypical hallucinations in psychotic and affective disorders and associations with non-benzodiazepine hypnotic use: the South London and Maudsley Case Register. *BMJ Open* **8**, e025216.

Bjelland I, Lie SA, Dahl AA, Mykletun A, Stordal E and Kraemer HC (2009) A dimensional versus a categorical approach to diagnosis: anxiety and depression in the HUNT 2 study. *International Journal of Methods in Psychiatric Research* **18**, 128–137.

Blei DM (2012) Probabilistic topic models. *Communications of the ACM* **55**, 77–84.

Blei DM, Ng AY and Jordan MI (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research* **3**, 993–1022.

Can D, Marín RA, Georgiou PG, Imel ZE, Atkins DC and Narayanan SS (2016) 'It sounds like…': a natural language processing approach to detecting counselor reflections in motivational interviewing. *Journal of Counseling Psychology* **63**, 343–350.

Compton MT, Craw J and Rudisch BE (2006) Determinants of inpatient psychiatric length of stay in an urban county hospital. *The Psychiatric Quarterly* **77**, 173–188.

Dictionary.com, LLC (2019) *Thesaurus.com*. Long Beach, CA: Lexico Publishing Group.

Fok ML, Stewart R, Hayes RD and Moran P (2014) The impact of co-morbid personality disorder on use of psychiatric services and involuntary hospitalization in people with severe mental illness. *Social Psychiatry and Psychiatric Epidemiology* **49**, 1631–1640.

Grilo CM, McGlashan TH, Quinlan DM, Walker ML, Greenfeld D and Edell WS (1998) Frequency of personality disorders in Two Age cohorts of psychiatric inpatients. *American Journal of Psychiatry* **155**, 140–142.

Grün B and Hornik K (2018). Topicmodels, v0.2-8. Available at https://cran.rproject.org/web/packages/topicmodels/index.html.

Habermeyer B, De Gennaro H, Frizi RC, Roser P and Stulz N (2018) Factors associated with length of stay in a Swiss mental hospital. *The Psychiatric Quarterly* **89**, 667–674.

Haslam N, Holland E and Kuppens P (2012) Categories versus dimensions in personality and psychopathology: a quantitative review of taxometric research. *Psychological Medicine* **42**, 903–920.

Heslin KC, Elixhauser A and Steiner CA (2015). *Hospitalizations Involving Mental and Substance Use Disorders Among Adults, 2012. HCUP Statistical Brief #191*. Rockville, MD: Agency for Healthcare Research and Quality.

Huprich SK (2018) Personality pathology in primary care: ongoing needs for detection and intervention. *Journal of Clinical Psychology in Medical Settings* **25**, 43–54.

Jacobs R, Gutacker N, Mason A, Goddard M, Gravelle H, Kendrick T and Gilbody S (2015) Determinants of hospital length of stay for people with serious mental illness in England and implications for payment systems: a regression analysis. *BMC Health Services Research* **15**, 439.

Jiménez RE, Lam RM, Marot M and Delgado A (2004) Observed-predicted length of stay for an acute psychiatric department, as an indicator of inpatient care inefficiencies. Retrospective case-series study. *BMC Health Services Research* **4**, 4.

Keown P, Holloway F and Kuipers E (2005) The impact of severe mental illness, co-morbid personality disorders and demographic factors on psychiatric bed use. *Social Psychiatry and Psychiatric Epidemiology* **40**, 42–49.

Kjelsås E and Augestad LB (2004) Gender, eating behavior, and personality characteristics in physically active students. *Scandinavian Journal of Medicine & Science in Sports* **14**, 258–268.

Kotov R, Krueger RF, Watson D, Achenbach TM, Althoff RR, Bagby RM, Brown TA, Carpenter WT, Caspi A, Clark LA, Eaton NR, Forbes MK, Forbush KT, Goldberg D, Hasin D, Hyman SE, Ivanova MY, Lynam DR, Markon K, Miller JD, Moffitt TE, Morey LC, Mullins-Sweatt SN, Ormel J, Patrick CJ, Regier DA, Rescorla L, Ruggero CJ, Samuel DB, Sellbom M, Simms LJ, Skodol AE, Slade T, South SC, Tackett JL, Waldman ID, Waszczuk MA, Widiger TA, Wright AGC and Zimmerman M (2017) The Hierarchical Taxonomy of Psychopathology (HiTOP): a dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology* **126**, 454–477.

Krismayer T, Schedl M, Knees P and Rabiser R (2019) Predicting user demographics from music listening information. *Multimedia Tools and Applications* **78**, 2897–2920.

Krueger RF, Derringer J, Markon KE, Watson D and Skodol AE (2012) Initial construction of a maladaptive personality trait model and inventory for DSM-5. *Psychological Medicine* **42**, 1879–1890.

Leontieva L and Gregory R (2013) Characteristics of patients with borderline personality disorder in a state psychiatric hospital. *Journal of Personality Disorders* **27**, 222–232.

Lugo V, de Oliveira SES, Hessel CR, Monteiro RT, Pasche NL, Pavan G, Motta LS, Pacheco MA and Spanemberg L (2019) Evaluation of DSM-5 and ICD-11 personality traits using the Personality Inventory for DSM-5 (PID-5) in a Brazilian sample of psychiatric inpatients. *Personality and Mental Health* **13**, 24–39.

Lynn R and Martin T (1997) Gender differences in extraversion, neuroticism, and psychoticism in 37 nations. *The Journal of Social Psychology* **137**, 369–373.

Manning CD and Schütze H (1999) *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

McCallum AK (2002). 'MALLET: A Machine Learning for Language Toolkit'. Available at http://mallet.cs.umass.edu.

McCoy Jr TH, Castro VM, Roberson AM, Snapper LA and Perlis RH (2016) Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. *JAMA Psychiatry* **73**, 1064–1071.

McCoy TH, Castro VM, Snapper LA, Hart KH, Januzzi JL, Huffman JC and Perlis RH (2017) Polygenic loading for major depression is associated with specific medical comorbidity. *Translational Psychiatry* **7**, e1238.

McCoy TH, Yu S, Hart KL, Castro VM, Brown HE, Rosenquist JN, Doyle AE, Vuijk PJ, Cai T and Perlis RH (2018) High throughput phenotyping for dimensional psychopathology in electronic health records. *Biological Psychiatry* **83**, 997–1004.

McLay RN, Daylo A and Hammer PS (2005) Predictors of length of stay in a psychiatric ward serving active duty military and civilian patients. *Military Medicine* **170**, 219–222.

Morey LC, Shea MT, Markowitz JC, Stout RL, Hopwood CJ, Gunderson JG, Grilo CM, McGlashan TH, Yen S, Sanislow CA and Skodol AE (2010) State effects of Major depression on the assessment of personality and personality disorder. *The American Journal of Psychiatry* **167**, 528–535.

Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S and Kohane I (2010) Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association: JAMIA* **17**, 124–130.

Newman L, Harris V, Evans LJ and Beck A (2018) Factors associated with length of stay in psychiatric inpatient services in London, UK. *The Psychiatric Quarterly* **89**, 33–43.

Newton-Howes G, Tyrer P, Anagnostakis K, Cooper S, Bowden-Jones O and Weaver T, COSMIC study team (2010) The prevalence of personality disorder, its comorbidity with mental state disorders, and its clinical significance in community mental health teams. *Social Psychiatry and Psychiatric Epidemiology* **45**, 453–460.

Pauselli L, Verdolini N, Bernardini F, Compton MT and Quartesan R (2017) Predictors of length of stay in an inpatient psychiatric unit of a general hospital in Perugia, Italy. *The Psychiatric Quarterly* **88**, 129–140.

Piccinelli M, Bortolaso P, Bolla E and Cioffi I (2016) Typologies of psychiatric admissions and length of inpatient stay in Italy. *International Journal of Psychiatry in Clinical Practice* **20**, 116–120.

Quirk SE, Berk M, Chanen AM, Koivumaa-Honkanen H, Brennan-Olsen SL, Pasco JA and Williams LJ (2016) Population prevalence of personality disorder and associations with physical health comorbidities and health care service utilization: a review. *Personality Disorders: Theory, Research, and Treatment* **7**, 136–146.

Řehůřek R and Sojka P (2010) Software framework for topic modelling with large corpora. In *Proceedings of LREC* 2010 *Workshop New Challenges for NLP Frameworks*. p. 46–50, 5 pp. ISBN 2-9517408-6-7.

Sevilla-Llewellyn-Jones J, Cano-Domínguez P, de-Luis-Matilla A, Peñuelas-Calvo I, Espina-Eizaguirre A, Moreno-Kustner B and Ochoa S (2017) Personality traits and psychotic symptoms in recent onset of psychosis patients. *Comprehensive Psychiatry* **74**, 109–117.

Skodol AE (2012) Personality disorders in DSM-5. *Annual Review of Clinical Psychology* **8**, 317–344.

Skodol AE (2018) Can personality disorders be redefined in personality trait terms? *American Journal of Psychiatry* **175**, 590–592.

Soeteman DI, Hakkaart-Van Roijen L, Verheul R and Van Busschbach J (2008) The economic burden of personality disorders in mental health care. *Journal of Clinical Psychiatry* **69**, 259–265.

Stevenson J, Datyner A, Boyce P and Brodaty H (2011) The effect of age on prevalence, type and diagnosis of personality disorder in psychiatric inpatients. *International Journal of Geriatric Psychiatry* **26**, 981–987.

Twomey CD, Baldwin DS, Hopfe M and Cieza A (2015) A systematic review of the predictors of health service utilisation by adults with mental disorders in the UK. *BMJ Open* **5**, e007575.

Tyrer P and Simmonds S (2003) Treatment models for those with severe mental illness and comorbid personality disorder. *The British Journal of Psychiatry. Supplement* **44**, S15–S18.

Tyrer P, Reed GM and Crawford MJ (2015) Classification, assessment, prevalence, and effect of personality disorder. *The Lancet* **385**, 717–726.

von Gunten A, Pocnet C and Rossier J (2009) The impact of personality characteristics on the clinical expression in neurodegenerative disorders – a review. *Brain Research Bulletin* **80**, 179–191.

Widiger TA (2011) Personality and psychopathology. *World Psychiatry* **10**, 103–106.

Wright AGC and Simms LJ (2015) A metastructural model of mental disorders and pathological personality traits. *Psychological Medicine* **45**, 2309–2319.

Yim W-W, Yetisgen M, Harris WP and Kwan SW (2016) Natural language processing in oncology: a review. *JAMA Oncology* **2**, 797–804.

Yu S, Kumamaru KK, George E, Dunne RM, Bedayat A, Neykov M, Hunsaker AR, Dill KE, Cai T and Rybicki FJ (2014) Classification of CT pulmonary angiography reports by presence, chronicity, and location of pulmonary embolism with natural language processing. *Journal of Biomedical Informatics* **52**, 386–393.

Zimmerman M (2016) Improving the recognition of borderline personality disorder in a bipolar world. *Journal of Personality Disorders* **30**, 320–335.

Zimmerman M and Morgan TA (2013) The relationship between bordersline personality disorder and bipolar disorder. *Dialogues in Clinical Neuroscience* **15**, 155–169.

Zimmerman M, Rothschild L and Chelminski I (2005) The prevalence of DSM-IV personality disorders in psychiatric outpatients. *The American Journal of Psychiatry* **162**, 1911–1918.

Zimmerman M, Chelminski I and Young D (2008) The frequency of personality disorders in psychiatric patients. *Psychiatric Clinics of North America* **31**, 405–420.