Biological Psychiatry

High Throughput Phenotyping for Dimensional Psychopathology in Electronic Health Records

Thomas H. McCoy Jr., Sheng Yu, Kamber L. Hart, Victor M. Castro, Hannah E. Brown, James N. Rosenquist, Alysa E. Doyle, Pieter J. Vuijk, Tianxi Cai, and Roy H. Perlis

ABSTRACT

BACKGROUND: Relying on diagnostic categories of neuropsychiatric illness obscures the complexity of these disorders. Capturing multiple dimensional measures of neuropathology could facilitate the clinical and neurobiological investigation of cognitive and behavioral phenotypes.

METHODS: We developed a natural language processing–based approach to extract five symptom dimensions, based on the National Institute of Mental Health Research Domain Criteria definitions, from narrative clinical notes. Estimates of Research Domain Criteria loading were derived from a cohort of 3619 individuals with 4623 hospital admissions. We applied this tool to a large corpus of psychiatric inpatient admission and discharge notes (2010–2015), and using the same cohort we examined face validity, predictive validity, and convergent validity with gold standard annotations.

RESULTS: In mixed-effect models adjusted for sociodemographic and clinical features, greater negative and positive symptom domains were associated with a shorter length of stay ($\beta = -.88$, $p = .001$ and $\beta = -1.22$, $p < .001$, respectively), while greater social and arousal domain scores were associated with a longer length of stay ($\beta = .93$, $p < .001$ and $\beta = .81$, $p = .007$, respectively). In fully adjusted Cox regression models, a greater positive domain score at discharge was also associated with a significant increase in readmission risk (hazard ratio = 1.22, $p < .001$). Positive and negative valence domains were correlated with expert annotation (by analysis of variance [$df = 3$], $R^2 = .13$ and .19, respectively). Likewise, in a subset of patients, neurocognitive testing was correlated with cognitive performance scores ($p < .008$ for three of six measures).

CONCLUSIONS: This shows that natural language processing can be used to efficiently and transparently score clinical notes in terms of cognitive and psychopathologic domains.

Keywords: Computed phenotype, Electronic health record, Natural language processing, Research Domain Criteria, Topic modeling, Transdiagnostic

https://doi.org/10.1016/j.biopsych.2018.01.011

The limitations of a categorical diagnostic system in neuropsychiatric illness have become increasingly apparent in an era of genomic study. A diagnostic category such as major depressive disorder (MDD) captures a large heterogeneous range of presentations (1). Co-occurrence of psychiatric disorders is the norm, conflating true comorbidity with different manifestations of the same underlying pathology, such as in cases of bipolar disorder (BPD) and anxiety disorders (2). The overlap in presentations and symptoms between disorders is not well captured—for example, this limitation manifests in the complexity of the relationship between mood disorders and psychotic disorders.

The information loss from categorization has become even more striking with the emergence of alternative means of defining the relationship between disorders. Twin and family studies dating back decades illustrated that while individual disorders are familial and heritable, an abundance of data now demonstrate the continuity between psychiatric disorders in terms of genomic liability and environmental risk (3–5).

Investigators frequently encounter the limitations of this system, and increasing attention has turned to multidimensional alternatives (6). The National Institute of Mental Health (NIMH) introduced the Research Domain Criteria (RDoC) as an alternative nosology focusing on linking clinical symptoms to relevant biology (7). These five domains—negative and positive valence, social function, cognition, and arousal—are intended to capture the full range of brain-associated function (8). Despite the appeal of RDoC as a means of facilitating translational studies, efficient assessment of these domains in clinical samples has yet to be established; it is intended as a research framework, not a clinical assessment per se. NIMH leadership has suggested that approaches incorporating "big data," or large clinical data sets, will be necessary for continued progress in understanding dimensional psychopathology (9–11). Still, the ability to estimate manifestation of these domains—even coarsely—in clinical data could greatly facilitate targeted investigations.

ISSN: 0006-3223

Natural language processing (NLP) refers to a broad set of methods extracting concepts or structured information from text (e.g., narrative clinical notes). These methods range from simple (e.g., matching particular strings in a block of text, or treating a document as a "bag of words") to extremely complex, incorporating context and attempting to extract meaning (12,13). In a clinical context, NLP provides a means of investigating phenotypic hypotheses not addressed by structured clinical data (e.g., health billing information or rating scales) (14). In psychiatry, diverse applications of NLP include identifying the presence or absence of depression in any given clinical visit and efforts to identify negative symptoms in psychosis, facilitating measures of the quantity of symptoms that are present (15–17). The utility of NLP has also been demonstrated outside of psychiatry, including the effective identification of the presence or absence of pulmonary embolism in radiology reports (18). Importantly, these are examples of restructuring text or identifying an individual symptom or outcome that could conceptually have been collected as structured data during the initial encounter. These examples apply NLP as a "force multiplier" by training models on expert annotations and then generalizing to many new cases in a supervised learning paradigm. In both cases—restructuring and supervised learning—a priori knowledge of a gold standard is assumed.

An alternative and complementary approach uses NLP to characterize notes without the assumption of known gold standard labels. Such methods assist in identifying unlabeled latent traits that are not yet well studied. We previously demonstrated the feasibility of applying NLP to extract multiple continuous symptom domains from psychiatric notes and found that the extracted dimensions improved the prediction of hospital readmission (19). However, this approach had two major limitations preventing broader application. First, it did not allow for inspection of the contributors to domain estimates and thus was not conducive to hypothesis generation. Second, it was computationally intensive and technically difficult to implement across health systems. Finally, the model used cohort-level score normalization that precluded online scoring. An ideal method would allow high throughput online estimates of existing clinical text, yield estimates with predictive and face validity, and allow the source of those estimates to be inspected. We describe a novel method for identifying estimates of loading for each of the five RDoC domains, distinct from our previous work with improved inspectability, portability, and performance. We demonstrate that this method has strong face validity and interpretability and that it improves the prediction of clinical outcomes compared with structured data alone.

METHODS AND MATERIALS

Overview and Data Set Generation

Sociodemographic and clinical data were extracted from the longitudinal electronic health record (EHR) of the Massachusetts General Hospital. Clinical data include billing (claims) codes, medication e-prescriptions, and narrative clinical notes. We included any individuals 18 years of age or older with between one and 10 inpatient psychiatric hospitalizations between 2010 and 2015. We determined principal clinical diagnoses based on the ICD-9 code at admission, incorporating any psychiatric diagnosis with at least 20 individuals represented in the cohort. These included schizophrenia (ICD-9 295.x, except 295.7), schizoaffective disorder (295.7), posttraumatic stress disorder (309.8), anxiety disorders (300.0/1/2), substance use disorders (291 or 292), psychosis not otherwise specified (298.9), MDD (296.2 or 296.3), BPD–manic (296.0/1/4), other BPD (296.5/6/7/8), and suicidality without other primary diagnosis (V628).

A datamart containing all clinical data was generated with the i2b2 server software (version 1.6; i2b2, Boston, MA), a computational framework for managing human health data (20–22). The Partners Institutional Review Board approved the study protocol, waiving the requirement for informed consent as detailed by 45 CFR §46.116.

Study Design and Analysis

Primary analyses used a cohort design with all patients admitted during the period noted above. No individuals were missing. The admission and discharge documentation were used to estimate RDoC domain scores at both time points for all encounters. In addition, clinical outcomes, including length of stay and psychiatric hospital readmission, were used to validate the clinical utility of the scores. Length of stay was defined as the discharge date minus the admission date. Psychiatric hospital readmission was defined as a second psychiatric hospitalization at Massachusetts General Hospital within 1 year (a period during which individuals would be highly likely to be readmitted to the index hospital).

Derivation of Estimated Research Domain Criteria Token List

The goal of subsequent steps in phenotype derivation was to derive a set of tokens (i.e., single words or sets of two words [bigrams]) reflecting individual RDoC domains in narrative notes. We developed a multistep process that used the text of DSM-IV-TR, a list of 10 to 50 seed unigrams or bigrams manually curated per domain based on expert consensus (THM, RHP) review of the RDoC workgroup statements, and psychiatric discharge summaries to identify terms that may be conceptually similar to those experts associate with each of the five RDoC domains (23); for an overview of the entire process, see Supplemental Figure S1. Both the DSM-IV and the corpus of narrative discharge notes were normalized using the Unified Medical Language System Lexical Variant Generation package (24). The corpus of narrative discharge notes was tokenized to unigrams and bigrams, and stop words were eliminated.

For subsequent steps, thresholding choices were made by inspection of the individual distribution based on the authors' experience with health record NLP method development (25). Choices to trim distributions were based on balancing the computational complexity of the task and breadth of symptoms captured, aiming to minimize overfitting risk to maximize portability. All thresholding choices were made before analysis of outcomes and were blind to token.

The DSM-IV-TR was then similarly preprocessed to generate unigram and bigram counts. DSM-IV-TR tokens were limited to those appearing in the narrative note corpus and further limited to unigrams occurring between 0.1% and 99% of the time and bigrams occurring four or more times. The

retained DSM tokens were weighted by inverse document frequency, with paragraphs treated as "documents." We then applied latent semantic analysis to the weighted paragraph-wise count data, using singular value decomposition to transform the token–paragraph association to token–topic association (26). Based on inspection of distribution, the top 300 topics were retained for the subsequent similarity analysis among words. For each seed word, we identified 50 unigrams and bigrams with the greatest cosine similarity in the DSM as that seed's candidate synonyms.

Next, the candidate synonyms for each RDoC domain were filtered to ensure that only synonyms associated with a domain seed term could appear in the final model. For each domain, we jointly tested the association between each seed word token and each candidate synonym token. Significant associations were identified as those with p values lower than a threshold chosen to control a false discovery rate of 10% (27). Candidate synonyms identified in the DSM were dropped if they were not associated with any curated seed word in the clinical corpus.

To further filter these candidate terms, we required occurrences of candidate terms to be predictive of occurrences of seed words in the clinical notes. Only seed words appearing in 10% to 90% of notes were considered, and candidate terms were limited to those appearing in 5% to 90% of the notes. We first performed a univariate screening and retained terms that were correlated with the token sum of the domain-specific seed terms with absolute rank correlation of $\geq$0.05. Subsequently, for each seed term, we subsampled 500 notes and fitted adaptive least absolute shrinkage and selection operator (LASSO) penalized logistic regression models to predict its occurrence in each note with features being tokens of the candidate terms that passed the univariate screening. Candidate words with zero coefficients are considered non-informative. We repeated this process 20 times for each seed term of each domain by randomly subsampling the 500 notes. For each domain, we then summarized the predictiveness of each candidate term based on the proportion of the times the term received a nonzero coefficient averaged over the fitting and the seed terms. The terms with nonzero frequency $>$5% were considered as the final set of synonyms for each domain. This process generated a set of tokens believed, by virtue of their derivation, to be associated with domain symptomatology.

Alternative to LASSO for Synonym Identification

While LASSO remains a widely used method for variable selection, it may fail to identify pairs of terms that overlap in their variance explained and thus may be less sensitive than alternatives (28,29). As a sensitivity analysis of the primary LASSO selection algorithm, we applied probabilistic topic modeling. Consistent with previous work, we applied latent Dirichlet allocation to fit a 75-topic model to the full note corpus (30). We then selected one topic per domain from this model by inspecting the posterior distribution to identify the topic under which the original seed words were most likely (e.g., the negative valence topic was selected as the topic under which the negative valence seed words were most probable). To build the final token list, we cut the five selected topic distributions at 95% of the cumulative probability distribution. The resulting token lists were then used to generate domain scores as described for the primary analysis. In the interest of providing the most portable scoring system (developed below), both token identification systems used the common "bag of words" model without using extensions such as negation, temporality detection, or word sense disambiguation.

Implementation of Scoring System

Having derived token lists associated with each domain, we then implemented a greatly simplified scoring code to facilitate dissemination, reapplication, and replication. The simplification involved first inverting the lexical variant index such that all lexical variants were added to the token list to be identified, then implementing basic preprocessing and tokenization code depending only on the Python standard library.

To assign an estimated domain score to each note, we used the percent of domain-specific terms appearing in the note. For example, if a hypothetical domain contained 10 terms in its final term list, a note containing five of those terms would be assigned a domain score of 5 of 100. In presentation of results, we multiply these values by 10 to facilitate readability. All notes for the full cohort (all psychiatric discharges between 2010 and 2015) were scored.

In sum, this method should be viewed in two distinct parts: the derivation of the words and phrases of interest, and then development of an easy-to-use scoring system based on the derived words and phrases. Token lists and the code used are available online at https://github.com/thmccoy/CQH-Dimensional-Phenotyper.

Validation Analysis

The optimal clinical RDoC assessment is not yet available, and therefore it is not possible to compare the NLP-based RDoC domain scores to "gold standard" evaluations. Instead, to validate the clinical utility of the scores, we adopted multiple convergent strategies to estimate validity. We first examined the predictive validity of the individual psychopathology estimates by associating these scores with various clinical outcomes for all psychiatric discharges between 2010 and 2015, adjusting for age, gender, public versus private insurance, and self-report of race/ethnicity and baseline clinical features. Specifically, we assessed the extent to which these domain estimates predict length of hospital stay (based on admission notes) and readmission (based on discharge notes) in regression models. We used as a metric of improvement the comparison to two nested models (likelihood ratio test). For length of hospital stay, we made domain estimates from the admission note; for individuals with multiple admissions, we accounted for multiple observations per subjects using mixed-effect models as a means of maximizing efficiency of analysis. For time to hospital readmission, we used Cox clustered regression with results censored at death or at 3 years.

Second, we examined face validity by plotting selected DSM-IV diagnoses in terms of symptom domains. Qualitatively, we also generated word clouds to allow for visualization of the terms loading onto each domain. To enhance the extent to which this word cloud reflects the contribution of individual words to predictions, we weighted this word cloud using the Gini Index from a random forest trained on individual token counts at discharge against readmission.

Table 1. Cohort Characteristics

Demographics	*N* = 3619
Age, Years, Mean (SD)	43.89 (16.62)
Gender, Male, *n* (%)	1840 (50.8)
Public Insurance, *n* (%)	2095 (57.9)
Emergency Department Admission, *n* (%)	2429 (67.1)
Race/Ethnicity, *n* (%)	
Asian	145 (4.0)
Black	343 (9.5)
Hispanic	315 (8.7)
Other	211 (5.8)
White	2605 (72.0)
Diagnosis at Admission, *n* (%)	
Major depressive disorder	774 (41.6)
Bipolar disorder, depressed	171 (9.2)
Bipolar disorder, manic	144 (7.7)
Psychosis not otherwise specified	272 (14.6)
Schizoaffective disorder	135 (7.3)
Schizophrenia	160 (8.6)
Comorbidity	
Log Charlson score, mean (SD)	0.61 (0.74)

Third, we examined convergent validity for two domains (negative and positive valence) by comparing scores to expert annotation of 200 randomly selected individuals drawn from the full discharge cohort. We used a set of anchor points developed by the authors (RHP, THM, HEB, JNR) for an American Medical Informatics Association NLP challenge corresponding to clinically meaningful anchor points, where 0 indicates an absence of symptoms, 1 indicates subthreshold symptoms or symptoms of questionable importance, 2 indicates threshold symptoms requiring outpatient treatment, and 3 indicates threshold symptoms that are likely to require inpatient treatment or emergent intervention (31). These anchor points were used by an expert clinician (RHP) familiar with the NIMH workgroup statements to score current symptoms reflecting negative and positive valence. The rater was blinded to computed scores and did not access token lists before scoring. Correlations were computed between estimated scores and expert rating.

Pilot Study of Neurocognitive Measures

As noted, the extent to which specific measures load onto specific domains has been asserted but not tested systematically. To demonstrate the potential application of automated scoring and future validation, we analyzed data from a battery of measures from the Cambridge Neuropsychological Test Automated Battery, collected within a random subset of individuals with psychiatric discharges between 2010 and 2015 during a systematic assessment for a cellular biobanking study (32–34). Pearson product moment correlations were examined between estimated cognition score and six measures of cognitive domains, with $p < .05$, $p < .06$, or $p < .008$ conservatively considered a corrected threshold for association. All analyses used R software (35).

In all cases, validation studies used either the full discharge cohort (predictive validity and face validity) or a subset (convergent validity). For an example of portability via application in a distinct cohort, see McCoy *et al.* (36).

RESULTS

We identified 3619 individuals with 4623 hospital discharges between 2010 and 2015, and sociodemographic and clinical descriptors are shown in Table 1. Figure 1 illustrates the distribution of cognition and negative valence estimated RDoC scores for individuals with MDD or BPD mania, at hospital admission (top) and discharge (bottom). While depressive symptoms are generally more severe among patients admitted for MDD, the range of depressive symptoms among those with mania illustrates the spectrum of mixed features. At discharge, depressive symptoms diminished in both groups, but cognitive symptoms did not change appreciably. To illustrate face

Figure 1. Domain comparison contour plots showing change between admission (top) and discharge (bottom). BPAD-M, bipolar disorder–mania; MDD, major depressive disorder.

Table 2. Length of Stay Regression Model With and Without Domain Scores From Admission Documentation

	Model With RDoC Domains		Model Without RDoC Domains	
	β (95% CI)	*p* Value	β (95% CI)	*p* Value
Negative	−.88 (−1.41 to −0.36)	.001[a]		
Positive	−1.22 (−1.61 to −0.84)	<.001[a]		
Cognitive	.47 (−0.17 to 1.11)	.154		
Social	.93 (0.40 to 1.46)	<.001[a]		
Arousal and Regulatory	.81 (0.22 to 1.40)	.007[a]		
Age, Years	.09 (0.07 to 0.11)	<.001[a]	.10 (0.08 to 0.12)	<.001[a]
Gender, Male	−.49 (−1.03 to 0.04)	.071	−.57 (−1.11 to −0.03)	.037[a]
Race, White	.20 (−0.41 to 0.82)	.520	−.39 (−1.00 to 0.23)	.218
Public Insurance	.05 (−0.48 to 0.59)	.844	.16 (−0.38 to 0.71)	.557
Log Charlson Score	−.36 (−0.80 to 0.07)	.104	−.47 (−0.91 to −0.03)	.038[a]

CI, confidence interval; RDoC, Research Domain Criteria.
[a]$p < .05$.

validity, Supplemental Figure S2 depicts individual terms contributing to the negative valence and cognitive RDoC domain estimates at discharge as word clouds.

Using the full cohort, we examined the association between individual domains extracted from hospital admission notes and length of hospital stay (mean ± SD, 9.67 ± 8.92 days). Table 2 (right) reports a mixed-effect model including only sociodemographic and coded clinical features, including diagnosis; Table 2 (left) adds the five domain scores (sensitivity analysis in Supplemental Figure S3). Among the individual domains, greater negative and positive symptom domains were associated with a shorter length of stay ($\beta = -.88$, $p = .001$ and $\beta = -1.22$, $p < .001$, respectively) while greater social and arousal domain scores were associated with a longer length of stay ($\beta = .93$, $p < .001$ and $\beta = .81$, $p < .007$, respectively). Adding domain scores improved model fit (likelihood ratio $\chi^2_5 = 107.12$, $p < 2.2 \times 10^{-16}$).

Next, we examined the association between domain scores in discharge notes and time to index hospital readmission, again using the full discharge cohort with 10,187 years of follow-up (median follow-up, 951 days). Once again, we compared a Cox regression model incorporating only coded sociodemographic and clinical data (Table 3; sensitivity analysis in Supplemental Figure S4). Greater positive domain score at discharge was associated with significant increase in readmission risk (hazard ratio = 1.22, $p < .001$) (i.e., for every 1 SD increase in positive valence score, the readmission hazard increased by 22%). Figure 2 illustrates Kaplan-Meier survival curves for time to readmission split by the median discharge positive valence estimate ($p < .001$).

Validation Against Expert Annotation

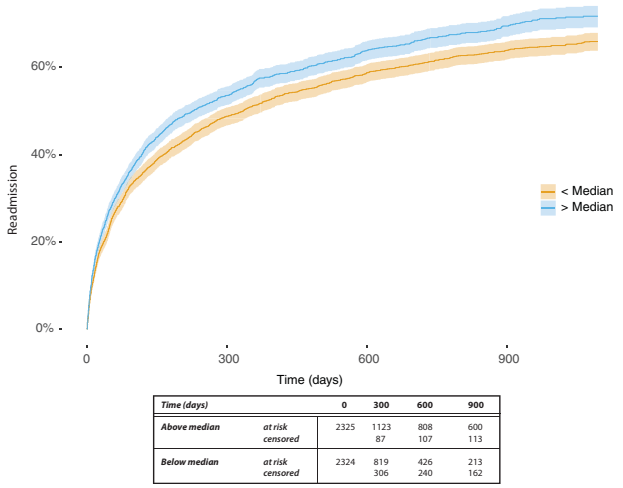
In a subset of 200 randomly selected individuals from this cohort, we examined the extent to which automated assignment of positive and negative valence severity correlated with a single expert annotator. Figure 3 depicts mean automated score, by expert annotation, for positive (top) and negative (bottom) valence for these 200 admissions. Positive and negative valence domains were correlated with expert annotation by analysis of variance ($R^2 = .13$ and .19, respectively).

Finally, as a pilot validation of the cognitive domain score we examined the association between estimated cognitive score and neurocognitive measures from the Cambridge Neuropsychological Test Automated Battery in a convenience sample of 11 individuals from the original cohort undergoing assessment as outpatients. Among the six measures, adjusting a priori for age and gender, we observed three associated

Table 3. Cox Regression of Time to Readmission With and Without Domain Scores From Discharge Documentation

Domain or Demographic Information	Model With RDoC Domains		Model Without RDoC Domains	
	HR (95% CI)	*p* Value	HR (95% CI)	*p* Value
Negative	0.98 (0.89–1.07)	.60		
Positive	1.22 (1.14–1.30)	3.34^{e-09}[a]		
Cognitive	0.96 (0.88–1.04)	.31		
Social	1.02 (0.95–1.11)	.59		
Arousal and Regulatory	0.90 (0.82–0.99)	.03[a]		
Age, Years	0.99 (0.99–0.99)	7.83^{e-11}[a]	0.99 (0.99–0.99)	8.49^{e-13}[a]
Gender, Male	0.97 (0.89–1.05)	.42	0.99 (0.92–1.08)	.90
Race, White	1.20 (1.09–1.33)	.0003[a]	1.22 (1.11–1.35)	4.92^{e-05}[a]
Public Insurance	1.44 (1.32–1.57)	5.55^{e-16}[a]	1.44 (1.32–1.57)	3.33^{e-16}[a]
Log Charlson Score	1.53 (1.43–1.63)	$<2.00^{e-16}$[a]	1.53 (1.43–1.63)	$<2.00^{e-16}$[a]

CI, confidence interval; HR, hazard ratio; RDoC, Research Domain Criteria.
[a]$p < .05$.



Time (days)		0	300	600	900
Above median	at risk	2325	1123	808	600
	censored		87	107	113
Below median	at risk	2324	819	426	213
	censored		306	240	162

Figure 2. Kaplan-Meier for time to readmission by split by median estimated Research Domain Criteria positive valence.

with cognition at hospital disposition with $p < .008$ (Supplemental Table S1).

Sensitivity Analysis

As a further examination of the robustness of these effects, we applied an alternative method to expand the token lists based on topic modeling that does not rely on LASSO. Results were markedly similar to those presented for primary analysis, despite the substantial methodologic difference. For example, in regression models for length of stay or readmission hazard, coefficients changed by less than 10% in all cases (see Supplemental Results).

DISCUSSION

We characterized 3619 individuals discharged from a psychiatric hospital between 2010 and 2015 in terms of symptom dimensions based upon the NIMH RDoC framework. We show that using NLP to calculate symptom dimension scores improves meaningfully on structured data alone, through inspectable collections of unigrams and bigrams. This method analyzes the vast corpora of narrative clinical notes available in EHRs to study the dimensional nature and implications of brain disease. We further demonstrate this method's predictive validity—explaining significant variance in length of hospital stay and readmission risk—as well as face and convergent validity.

Standard applications of EHR data sets have drawn on methodologies built for analyses of health claims data dating back decades. While well established, these methods neglect a growing understanding of the complexity of psychopathology. They assume categorical diagnoses, even though such diagnoses are often unreliable and divorced from the underlying neurobiology, including burgeoning research showing an overlap between major psychiatric diseases (and some neurologic diseases) in terms of common genetic risk (3,4).

As products of routine clinical care, detailed narrative notes provide an opportunity to recapture the complexity of psychopathology but require the extraction of relevant symptoms

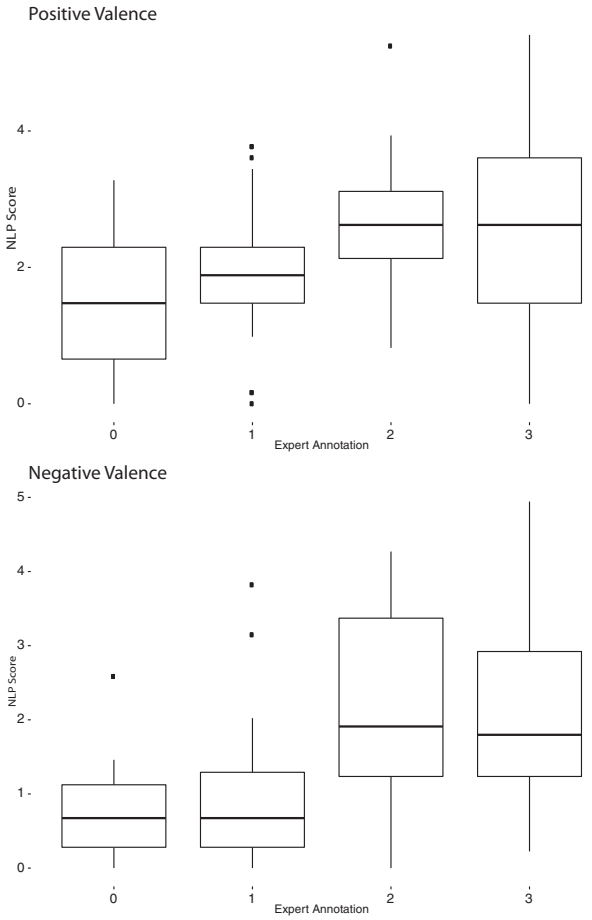


Figure 3. Comparison of automated natural language processing (NLP) score and expert annotation (positive top/negative bottom).

from unstructured data (14). Because structured elements are frequently incompletely coded, previous extraction work focused on validating classifiers for categorical phenotypes not coded at the time of diagnosis (15,16). However, an alternative approach that may track more closely with the biologically valid phenotypes is to extract multiple contiguous conceptual symptom domains rather than individual diagnostic labels (37). This approach differs fundamentally in methods and goals by operating independent of gold standard labels and estimates, instead aiding in creating such labels and estimates (19). In addition to offering new avenues for investigating causal biology, these multidimensional estimation-based approaches may be more scalable at the systems level. Instead of building a classifier per disease, multidimensional estimators could be used to create high-dimensionality-concept spaces within which clinically interesting categorical diagnoses are positioned.

In a previous investigation, we developed a dimensional phenotyping approach that demonstrated predictive validity in narrative clinical notes (19). While valuable as a proof-of-concept study, a lack of transparency coupled with computational complexity limited its application. To overcome these limitations, we report a phenotyping system based on multiple estimates instead of clinical categorization, and show the

distribution of these estimates of dimensional pathology across an inpatient psychiatric population. We demonstrate that these dimensions differ in face validity between diagnostic categories, while also illustrating substantial overlap. In addition, they help to explain length of hospital stay and readmission risk beyond the variance captured by available structured data.

The impact of positive valence score on readmission, which based upon inspection of token lists strongly reflects substance use disorders, appears to have face validity; we note the important distinction that the positive affect domain of RDoC reflects disorders of positive affect (e.g., reward), not positive affect per se. Substance abuse admissions to the psychiatric unit at this hospital are typically brief. Similarly, the association between negative valence and shorter length of stay may reflect acute stabilization of suicidal patients (i.e., those with high negative valence) who are discharged once risk diminishes. The replication of these results when applying an entirely different term-amplification (i.e., synonym-generation) method suggests their robustness.

We note several key limitations. First, this method will benefit from further validation against larger gold standard measures drawn from individual clinical assessment batteries. Such efforts are ongoing but await greater agreement on the necessary assessments and cohort generation. For two domains, positive and negative affect, we were able to draw on clinical expert annotations to examine convergent validity. The modest correlation here suggests opportunity to further refine concept detection as better gold standards are developed. For a third domain, cognition, we were able to draw on neuropsychological testing available in a small subset of the cohort to examine correlation between an estimate of cognitive symptomatology and objective cognitive measures. Validation of arousal and social domains may require additional data collection but is another important future direction.

Another limitation is that while this method is readily portable to other health data, the portability and generalizability of our method must be demonstrated, as with any such tool. In particular, it will be useful to examine the extent to which incorporating other note types (e.g., outpatient visit or inpatient progress notes) or the use of structured data from rating scales may improve the precision of estimating symptom domains or allow time series analysis. For an illustration of the application of this method to a biological question in a different patient population, see McCoy *et al.* (36). We present this work together with dependency-free open source software that is readily applicable to other narrative note sets to encourage replication efforts. Achieving this portability entailed constraints on toolchain and algorithmic complexity. For example, negative findings (e.g., "no evidence of psychosis") are common in medical documentation (38,39). Systems for automated detection of negation are an active area of research, and further work is needed to incorporate these findings in a fashion that preserves portability (40–42). Finally, we would emphasize that the approach we used constrains the symptom dimensions to correspond to those specified by NIMH workgroups; an important area of future work will be understanding the extent to which less constrained approaches to deriving symptom domains do or do not correspond to these hypothesized domains.

With these caveats in mind, the present report represents a key next step in developing and validating a simple, transparent, and scalable NLP approach to extracting dimensional psychopathology. Such strategies may ensure that EHRs can be leveraged for modern data-driven analysis without abandoning the wealth of dimensional data they contain.

ACKNOWLEDGMENTS AND DISCLOSURES

This work was supported by National Human Genome Research Institute Grant No. 1P50HG00738-01 (to RHP), National Institute of Mental Health Grant No. 1R01MH106577-01A1 (to RHP), and the Broad Institute Stanley Center Fellowship (to THM). The sponsors had no role in study design, writing of the report, or data collection, analysis, or interpretation. The corresponding and senior authors had full access to all data and made the decision to submit for publication.

We thank the National Institute of Mental Health Research Domain Criteria Unit members who contributed to manual review and curation of the Research Domain Criteria term list.

RHP reports grants from the National Human Genome Research Institute and grants from the National Institute of Mental Health during the conduct of this study. RHP received personal fees from Genomind, Healthrageous, Perfect Health, Psy Therapeutics, and RID Ventures. THM reports grants from the Broad Institute and Brain and Behavior Foundation. The other authors report no biomedical financial interests or potential conflicts of interest.

ARTICLE INFORMATION

From the Center for Quantitative Health and Department of Psychiatry (THM, KLH, VMC, HEB, JNR, AED, PJV, RHP), Massachusetts General Hospital and Harvard Medical School, and the Harvard School of Public Health (SY, TC), Boston, Massachusetts; and Tsinghua University (SY), Beijing, China.

THM and SY contributed equally to this work. TC and RHP contributed equally to this work.

Address correspondence to Thomas H. McCoy Jr., M.D., Massachusetts General Hospital, Simches Research Building, 6th floor, Boston, MA 02114; E-mail: thmccoy@partners.org.

Received Aug 10, 2017; revised Dec 15, 2017; accepted Jan 8, 2018.

Supplementary material cited in this article is available online at https://doi.org/10.1016/j.biopsych.2018.01.011.

REFERENCES

1. Zimmerman M, Ellison W, Young D, Chelminski I, Dalrymple K (2015): How many different ways do patients meet the diagnostic criteria for major depressive disorder? Compr Psychiatry 56:29–34.
2. Pavlova B, Perlis RH, Alda M, Uher R (2015): Lifetime prevalence of anxiety disorders in people with bipolar disorder: A systematic review and meta-analysis. Lancet Psychiatry 2:710–717.
3. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, *et al.* (2015): An atlas of genetic correlations across human diseases and traits. Nat Genet 47:1236–1241.
4. Cross-Disorder Group of the Psychiatric Genomics Consortium, Lee SH, Ripke S, Neale BM, Faraone SV, Purcell SM, *et al.* (2013): Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat Genet 45:984–994.
5. Gilman SE, Ni MY, Dunn EC, Breslau J, McLaughlin KA, Smoller JW, *et al.* (2015): Contributions of the social environment to first-onset and recurrent mania. Mol Psychiatry 20:329–336.
6. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, *et al.* (2010): Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. Am J Psychiatry 167:748–751.
7. Sanislow CA, Pine DS, Quinn KJ, Kozak MJ, Garvey MA, Heinssen RK, *et al.* (2010): Developing constructs for psychopathology research: Research domain criteria. J Abnorm Psychol 119:631–639.
8. Morris SE, Cuthbert BN (2012): Research Domain Criteria: Cognitive systems, neural circuits, and dimensions of behavior. Dialogues Clin Neurosci 14:29–37.

9. Gordon J (2017): The future of RDoC. National Institute of Mental Health. Available at: https://www.nimh.nih.gov/about/director/messages/2017/the-future-of-rdoc.shtml. Accessed January 21, 2018.
10. Insel TR, Cuthbert BN (2015): Brain disorders? Precisely. *Science*. 348: 499-500.
11. Redish AD, Gordon JA (2016): Computational psychiatry: New perspectives on mental illness. Cambridge, MA: MIT Press.
12. Nadkarni PM, Ohno-Machado L, Chapman WW (2011): Natural language processing: An introduction. J Am Med Inform Assoc 18: 544–551.
13. Manning CD, Schütze H (1999): Foundations of statistical natural language processing. Cambridge, MA: MIT Press.
14. Forbush TB, Gundlapalli AV, Palmer MN, Shen S, South BR, Divita G, *et al.* (2013): "Sitting on pins and needles": Characterization of symptom descriptions in clinical notes. AMIA Jt Summits Transl Sci Proc 2013:67–71.
15. Perlis RH, Iosifescu DV, Castro VM, Murphy SN, Gainer VS, Minnier J, *et al.* (2012): Using electronic medical records to enable large-scale studies in psychiatry: Treatment resistant depression as a model. Psychol Med 42:41–50.
16. Patel R, Jayatilleke N, Broadbent M, Chang CK, Foskett N, Gorrell G, *et al.* (2015): Negative symptoms in schizophrenia: A study in a large clinical sample of patients using a novel automated method. BMJ Open 5:e007619.
17. Gorrell G, Jackson R, Roberts A, Stewart R (2013): Finding negative symptoms of schizophrenia in patient records. In: Proceedings of the Workshop on NLP for Medicine and Biology associated with RANLP 2013, Hissar, Bulgaria, September 13, 2013, 9–17.
18. Yu S, Kumamaru KK, George E, Dunne RM, Bedayat A, Neykov M, *et al.* (2014): Classification of CT pulmonary angiography reports by presence, chronicity, and location of pulmonary embolism with natural language processing. J Biomed Inform 52:386–393.
19. McCoy TH, Castro VM, Rosenfield HR, Cagan A, Kohane IS, Perlis RH (2015): A clinical perspective on the relevance of research domain criteria in electronic health records. Am J Psychiatry 172:316–320.
20. Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, Phillips LC, *et al.* (2007): Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. AMIA Annu Symp Proc 548–552.
21. Murphy S, Churchill S, Bry L, Chueh H, Weiss S, Lazarus R, *et al.* (2009): Instrumenting the health care enterprise for discovery research in the genomic era. Genome Res 19:1675–1681.
22. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, *et al.* (2010): Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 17:124–130.
23. National Institute of Mental Health. Development of the RDoC framework. Available at: https://www.nimh.nih.gov/research-priorities/rdoc/development-of-the-rdoc-framework.shtml. Accessed January 21, 2018.
24. The SPECIALIST lexicon and lexical tools. In: UMLS reference manual. Bethesda, MD: National Library of Medicine
25. Yu S, Liao KP, Shaw SY, Gainer VS, Churchill SE, Szolovits P, *et al.* (2015): Toward high-throughput phenotyping: Unbiased automated feature extraction and selection from knowledge sources. J Am Med Inform Assoc 22:993–1000.
26. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990): Indexing by latent semantic analysis. J Am Soc Inf Sci 41:391.
27. Cai TT, Liu W (2016): Large-scale multiple testing of correlations. J Am Stat Assoc 111:229–240.
28. Austin PC, Tu JV (2004): Bootstrap methods for developing predictive models. Am Stat 58:131–137.
29. Zou H (2006): The adaptive Lasso and its oracle properties. J Am Stat Assoc 101:1418–1429.
30. McCoy TH, Castro VM, Snapper LA, Hart KL, Januzzi JL, Huffman JC, *et al.* (2017): Polygenic loading for major depression is associated with specific medical comorbidity. Translational Psychiatry 7:e1238.
31. Filannino M, Stubbs A, Uzuner Ü (2016): 2016 CEGS N-GRID shared-tasks and workshop on challenges in natural language processing for clinical data. Available at: https://www.i2b2.org/NLP/RDoCforPsychiatry/. Accessed January 21, 2018.
32. Falconer DW, Cleland J, Fielding S, Reid IC (2010): Using the Cambridge Neuropsychological Test Automated Battery (CANTAB) to assess the cognitive impact of electroconvulsive therapy on visual and visuospatial memory. Psychol Med 40:1017–1025.
33. Egerhazi A, Berecz R, Bartok E, Degrell I (2007): Automated Neuropsychological Test Battery (CANTAB) in mild cognitive impairment and in Alzheimer's disease. Prog Neuropsychopharmacol Biol Psychiatry 31:746–751.
34. Levaux MN, Potvin S, Sepehry AA, Sablier J, Mendrek A, Stip E (2007): Computerized assessment of cognition in schizophrenia: promises and pitfalls of CANTAB. Eur Psychiatry 22:104–115.
35. R Development Core Team (2016): R: A language and environment for statistical computing. 3.1.1 ed. Vienna, Austria: R Foundation for Statistical Computing.
36. McCoy TH Jr, Castro VM, Hart KL, Pellegrini AM, Yu S, Cai T, Perlis RH (2018): Genome-wide association study of dimensional psychopathology using electronic health records. Biol Psychiatry 83:1005–1011.
37. McCoy TH Jr, Castro VM, Roberson AM, Snapper LA, Perlis RH (2016): Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. JAMA Psychiatry 73:1064–1071.
38. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG (2001): Evaluation of negation phrases in narrative clinical reports. Proc AMIA Symp 105–109.
39. Harkema H, Dowling JN, Thornblade T, Chapman WW (2009): Context: An algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biomed Inform 42:839–851.
40. Wu S, Miller T, Masanz J, Coarr M, Halgrim S, Carrell D, *et al.* (2014): Negation's not solved: Generalizability versus optimizability in clinical natural language processing. PLoS One 9:e112774.
41. Mehrabi S, Krishnan A, Sohn S, Roch AM, Schmidt H, Kesterson J, *et al.* (2015): DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. J Biomed Inform 54:213–219.
42. Mukherjee P, Leroy G, Kauchak D, Rajanarayanan S, Romero Diaz DY, Yuan NP, *et al.* (2017): NegAIT: A new parser for medical text simplification using morphological, sentential and double negation. J Biomed Inform 69:55–62.