Published in final edited form as:

Am J Psychiatry. 2015 April ; 172(4): 316–320. doi:10.1176/appi.ajp.2014.14091177.

A Clinical Perspective on the Relevance of Research Domain Criteria in Electronic Health Records

Thomas H. McCoy, M.D.,

Victor M. Castro, M.S.,

Hannah R. Rosenfield, B.A.,

Andrew Cagan, B.S.,

Isaac S. Kohane, M.D.,

Roy H. Perlis, M.D., M.Sc.

Center for Experimental Drugs and Diagnostics, the Department of Psychiatry, and the Center for Human Genetic Research, Massachusetts General Hospital, Boston; and the Department of Medicine, Brigham and Women's Hospital, Boston.

Abstract

Objective: The limitations of the DSM nosology for capturing dimensionality and overlap in psychiatric syndromes, and its poor correspondence to underlying neurobiology, have been well established. The Research Domain Criteria (RDoC), a proposed dimensional model of psychopathology, may offer new insights into psychiatric illness. For psychiatric clinicians, however, because tools for capturing these domains in clinical practice have not yet been established, the relevance and means of transition from the categorical system of DSM-5 to the dimensional models of RDoC remains unclear. The authors explored a method of extracting these dimensions from existing electronic health record (EHR) notes.

Method: The authors used information retrieval and natural language processing methods to extract estimates of the RDoC dimensions in the EHRs of a large health system. They parsed and scored EHR documentation for 2,484 admissions covering 2,010 patients admitted to a psychiatric inpatient unit between 2011 and 2013. These domain scores were compared with DSM-IV-based ICD-9 codes to assess face validity. As a measure of predictive validity, these scores were examined for association with two outcomes: length of hospital stay and time to all-cause hospital readmission. Together, these analyses were intended to address the extent to which RDoC symptom domains might capture clinical features already available in narrative notes but not reflected in DSM diagnoses.

Results: In mixed-effects models, loadings for the RDoC cognitive and arousal domains were associated with length of hospital stay, while the negative valence and social domains were associated with hazard of all-cause hospital readmission.

Conclusions: These findings show that a computationally derived tool based on RDoC workgroup reports identifies symptom distributions in clinician notes beyond those captured by

Address correspondence to Dr. Perlis (rperlis@mgh.harvard.edu).

ICD-9 codes, and these domains have significant predictive validity. More generally, they point to the possibility that clinicians already document RDoC-relevant symptoms, albeit not in a quantified form.

The limitations of the modern psychiatric diagnostic system embodied in DSM-IV and DSM-5 are well established (1–3). They include an inability to capture the dimensional nature of many disorders, a reliance on combining heterogeneous symptoms or presentations, and a high rate of symptom overlap that may result in patients being diagnosed with multiple disorders simultaneously. Of perhaps greater concern, these diagnostic categories appear to correspond poorly to the neurobiology of psychiatric illness; for example, data from recent large-scale genome-wide association studies indicate marked overlap in liability across five major psychiatric disorders (4). Moreover, some symptom domains that are known to be important in multiple disorders, such as executive function, are simply not captured by DSM criteria (5).

The RDoC were proposed to facilitate a broader and more integrative phenotypic assessment of neuropsychiatric disease, corresponding more directly to underlying neurobiology (2, 6). The five symptoms domains—negative valence, positive valence, cognitive functioning, arousal, social processes—are described in the proceedings of workgroups convened by NIMH in 2011 and 2012 (http://www.nimh.nih.gov/research-priorities/rdoc/index.shtml). The institutional emphasis on this new nosology has been made explicit, with the NIMH director stating, “We plan to use RDoC as a framework for guiding our funding” (7).

This initiative poses two challenges for clinicians and clinical investigators in psychiatry. First, at present there is no consensus on the means of capturing these domains in clinical settings. Second, because these categories are constructed de novo to be independent of the DSM nosology, they effectively orphan decades of expertly collected clinical data. As a consequence, potentially powerful database studies of psychiatric illnesses or risk factors (8, 9) made newly feasible by electronic health records (EHRs) cannot make use of RDoC-informed analysis.

In an effort to preserve the potential contribution of past experience as captured in EHRs to future RDoC-informed investigation, we applied natural language processing and information retrieval techniques to calculate estimates of RDoC domain loading from clinical narrative notes. To this end, we scored narrative clinical free text from alarge New England health system according to the text’s similarity to documents collected by Internet searches on terms from the RDoC workgroup concept matrix. While this process represents only a starting point in translating between clinical and research measures, it provides evidence that clinicians do encode dimensional elements in their narrative notes that can be understood in the context of RDoC.

METHOD

Cohort and Outcome Derivation

We used the i2b2 server software, version 1.6.04 (10), to access and manipulate demographic data, billing codes, and narrative notes from the EHRs of a large Boston-based hospital that includes a network of primary and specialty care settings.

The cohort examined here was composed of 2,484 narrative admission notes from 2,010 patients admitted to an adult psychiatric inpatient unit at an academic medical center between 2011 and 2013. No inclusion or exclusion criteria were applied; all patients were age 18 or older at time of admission.

Outcome measures were those that could be derived directly from coded data available in the EHR as an artifact of routine care. Length of hospital stay was determined from billing data. Time to all-cause hospital readmission was determined by examining the subsequent 2-year period to identify readmissions to the index hospital. (In Massachusetts, readmissions would typically occur at the same hospital, but because it is an “open” system, we cannot exclude the possibility of interim admissions elsewhere.)

Development and Application of Scoring Tool

Methods for scoring individual notes are described in Figure 1. In summary, we began with individual terms listed in the RDoC workgroup documents and identified synonyms or other relevant concepts using an automated web search. Each note was then scored according to its similarity to the body of documents identified by the search, yielding five estimated RDoC-like domains. Thus, higher levels of a given domain refer to greater loading for terms in that domain; while in general this would indicate greater severity (e.g., for negative valence), for the social domain in particular, this could correspond to greater social involvement (a strength) or more social activity (which would be a concern in mania), with similar effects in arousal.

We note an important difference between this approach and traditional approaches to clinical text. More often, pre-defined query language (e.g., sets of terms describing smoking status or treatment resistance in depression) is applied to search clinical text. Here, we reverse the typical search process, applying the clinical text (narrative note) to query an external corpus (the set of documents returned by searching the Internet for terms described by the RDoC workgroup statements).

Analysis

All estimated RDoC (eRDoC) domain scores were z-transformed in order to aid in interpretability. For descriptive purposes, we first examined the distribution of scores across the subset of individuals diagnosed at admission with major depressive disorder (N=327), schizophrenia (N=80), or schizoaffective disorder (N=81). Analysis of variance with post hoc pairwise tests was used to contrast these three groups.

Next, associations between eRDoC domain scores, diagnoses, and length of hospital stay were examined using mixed-effects models, in order to account for multiple clustered

observations per patient, among all individuals. First, we fitted a model including demographic variables without these factors; then we added these factors, and the change in model fit was assessed by likelihood ratio test. A standard Bonferroni correction would require a p threshold of <0.01 for statistical significance with an experiment-wide alpha of 0.05.

Finally, time to all-cause hospital readmission after discharge used survival analysis with results censored at readmission, death, or 2 years, whichever came first. Effects of domain scores were examined using Cox proportional hazards models, adjusted for demographic and clinical variables. Analyses were conducted with Stata, version 13 (StataCorp, College Station, Tex.).

RESULTS

The demographic and clinical characteristics of the inpatient cohort are summarized in Table 1. Given that this was a psychiatric unit in a general hospital, the mean Charlson comorbidity index of 3.4 suggests substantial medical comorbidity. Notably, only about a quarter of individuals were diagnosed with major depressive disorder, schizophrenia, or schizoaffective disorder; other common diagnoses included mood disorder not otherwise specified and psychosis not otherwise specified, which are widely used as “placeholder” diagnoses, further illustrating the complexity and limitations of DSM-5 diagnosis.

Figure 2 illustrates the distribution, in each selected because we anticipated it would differ predictably between schizophrenia and major depressive disorder. By analysis of variance, negative valence was significantly greater in patients with schizoaffective disorder and major depressive disorder compared with those with schizophrenia (overall F=27.63; p<0.001; post hoc pairwise tests were significant for all groups [p<0.01]). On the other hand, Figure 2 also illustrates the substantial overlap between the groups on this dimensional measure.

Next, we examined the predictive validity of eRDoC domain scores. Table 2 summarizes results of a mixed-effects regression model examining the association between individual scores and length of hospital stay, adjusted for differences in demographic and clinical characteristics. The initial model incorporated only age, sex, race, insurance status, and medical comorbidity. Adding terms for the eRDoC domain scores significantly improved model fit (likelihood ratio χ^2=32.75, df=5, p<0.001). In the adjusted model, both the cognitive and arousal domains were significantly associated with length of hospital stay (Table 2). Coefficients changed modestly with the addition of ICD-9 diagnosis to the model: for the cognitive domain, the beta coefficient was 3.16 (p=0.003); for the arousal domain, the beta coefficient was −1.93 (p=0.075).

Finally, we examined the association between admission note eRDoC domain scores and time to all-cause hospital readmission after discharge using Cox regression models (see Table S1 in the data supplement that accompanies the online edition of this article): both the negative valence and social domains were significantly associated with readmission risk; for the negative valence domain, the hazard ratio was 1.95 (p=0.03), and for the social domain, it was 0.59 (p=0.04). Once again, adding ICD-9 diagnosis to the models did not change the

coefficients associated with negative and social domains. The association between negative valence score and readmission is illustrated by a Kaplan-Meier survival curve in Figure 3.

DISCUSSION

In this analysis of data from more than 2,000 patients, leveraging 53,285 documents encompassing more than 89,973,395 words, we demonstrated that eRDoC domain scores are associated with clinically meaningful outcomes, in a manner not fully accounted for by ICD-9 diagnosis code. This result should provide some reassurance to clinicians that their notes do contain relevant detail for deriving dimensional measures of illness: like Molière's Bourgeois Gentleman, speaking prose without knowing it, clinicians may already speak some RDoC.

We emphasize that these measures are not a substitute for assessment of RDoC domains in individual patients. A key next step before applying our approach more broadly would be clinical validation: comparing estimated domain scores with actual measures of severity using RDoC paradigms nominated by the workgroups. The availability of the software we describe here may facilitate such studies by allowing high-throughput screening of health system cohorts based on existing EHR data, followed by targeted recruitment for prospective validation studies. Before RDoC can truly be used in clinical practice, it will be necessary to derive clinical assessment tools that are relatively brief, are easily implemented and disseminated, and capture at least some of the predictive validity of traditional diagnostic systems. Our work provides some preliminary encouragement that, since clinical features corresponding to RDoC domains are already assessed to some extent in clinical practice, developing clinical tools may be feasible.

For now, clinicians might be best advised simply to be aware of the usefulness of dimensional models to capture psychopathology. Strategies that will render clinical reports more forward-compatible with RDoC include comprehensive psychiatric reviews of systems, characterization of individual symptoms, and, wherever possible, use of clinically valid quantitative measures of symptom domains, recognizing that few of these measures capture a single dimension. For example, while the Quick Inventory of Depressive Symptomatology–Self-Report is an efficient and useful means of measuring depressive symptoms, it samples the negative valence, positive valence, social, and even cognitive domains.

Despite these limitations, an advantage of this analysis is that it represents a general clinical population, not a cohort of healthy individuals or a specific ICD-9 diagnosis. This capacity allows us to quantify cross-disorder variation in symptom domains consistent with the intent of the RDoC initiative. To our knowledge, Figure 2 represents one of the first depictions of how symptoms across multiple diverse domains vary in cross-disorder cohorts—in essence, it provides a snapshot of severe psychopathology in an academic medical center.

Taken together, our results demonstrate the feasibility of applying large-scale scoring of clinical free text drawn from EHRs to characterize RDoC-like symptom dimensions in clinical populations. Clinicians are already accustomed to measuring and describing some

symptoms that load onto RDoC dimensions. By enabling EHR data to be used to study symptom dimensions, the approach we describe rescues the massive corpus of narrative notes that describe psychopathology over a decade or more and may help clinicians begin to think in terms of dimensional measures.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Dr. Perlis is supported by NIMH grant MH086026 and the Stanley Center for Psychiatric Research at the Broad Institute. Dr. McCoy is a participant in NIMH grant 5R25MH094612.

Dr. Perlis has served on advisory boards or served as a consultant for Genomind, Healthrageous, Perfect Health, Pfizer, Proteus Biomedical, PsyBrain, RID Ventures, and he receives royalties from Concordant Rater Systems. The other authors report no financial relationships with commercial interests.

REFERENCES

1. Casey BJ, Craddock N, Cuthbert BN, et al. : DSM-5 and RDoC: progress in psychiatry research? Nat Rev Neurosci 2013; 14:810–814 [PubMed: 24135697]
2. Insel T, Cuthbert B, Garvey M, et al. : Research Domain Criteria (RDoC): toward a new classification framework for research on mental disorders. Am J Psychiatry 2010; 167:748–751 [PubMed: 20595427]
3. Cuthbert BN, Insel TR: Toward the future of psychiatric diagnosis: the seven pillars of RDoC. BMC Med 2013; 11:126 [PubMed: 23672542]
4. Lee SH, Ripke S, Neale BM, et al. : Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat Genet 2013; 45:984–994 [PubMed: 23933821]
5. Barch DM, Keefe RS: Anticipating DSM-V: opportunities and challenges for cognition and psychosis. Schizophr Bull 2010; 36:43–47 [PubMed: 19923191]
6. Cuthbert BN: The RDoC framework: facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. World Psychiatry 2014; 13:28–35 [PubMed: 24497240]
7. Insel T: Director's Blog: Research Domain Criteria–RDoC. March 6, 2012. http://www.nimh.nih.gov/about/director/2012/research-domain-criteria-rdoc.shtml
8. Castro VM, Clements CC, Murphy SN, et al. : QT interval and anti-depressant use: a cross sectional study of electronic health records. BMJ 2013; 346:f288 [PubMed: 23360890]
9. Perlis RH, Iosifescu DV, Castro VM, et al. : Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. Psychol Med 2012; 42:41–50 [PubMed: 21682950]
10. Murphy SN, Mendis M, Hackett K, et al. : Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. AMIA Annu Symp Proc, Oct 11, 2007, pp 548–552 [PubMed: 18693896]
11. Zelikovitz S, Kogan M: Using web searches on important words to create background sets for LSI classification, in Proceedings of the 19th International Florida AI Research Society Conference. Palo Alto, Calif, American Association for Artificial Intelligence, 2006
12. Mattmann C, Zitting J: Tika in Action. Greenwich, Conn, Manning Publications, 2011
13. Kohlschütter C, Fankhauser P, Nejdl W: Boilerplate detection using shallow text features, in Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM 2010). New York, Association for Computing Machinery, 2010, pp 441–450
14. Kiss T, Strunk J: Unsupervised multilingual sentence boundary detection. Comput Linguist 2006; 32:485–525

15. Marcus MP, Santorini B, Marcinkiewicz MA: Building a large annotated corpus of English: the Penn Treebank. Comput Linguist 1993; 19:313–330
16. Miller GA, Beckwith R, Felbaum C, et al. : Introduction to WordNet: an on-line lexical database. International Journal of Lexicography. 1990; 3:235–244
17. Bird S, Klein E, Loper E: Natural Language Processing With Python. Sebastopol, Calif, O'Reilly Media, 2009
18. Salton G, Wong A, Yang CS: A vector space model for automatic indexing. Commun ACM 1975; 18:613–620
19. Bradford R: An empirical study of required dimensionality for large-scale latent semantic indexing applications, in Proceedings of the 17th ACM Conference on Information and Knowledge Management, 2008, pp 153–162
20. Deerwester S: Improving information retrieval with latent semantic indexing, in Proceedings of the 51st Annual Meeting of the American Society for Information Science 1988
21. Rehurek R: Subspace tracking for latent semantic analysis, in ECIR '11: Proceedings of the 33rd European Conference on Advances in Information Retrieval, 2011, pp 289–300
22. Singhal A: Modern information retrieval: a brief overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering. 2001; 24:35–43

CLINICAL QUESTION

A third-year resident has a 32-year-old patient whose 10-year history includes admissions for both schizophrenia and bipolar disorder. In the present episode, the patient presents with mood elevation and paranoid delusions. The family is confused and concerned about the patient's diagnosis, particularly because the two illnesses are said to have different clinical outcomes. The resident asks whether the patient's past clinical history could be reanalyzed through the perspective of the National Institute of Mental Health (NIMH) Research Domain Criteria (RDoC) to better predict the patient's course.

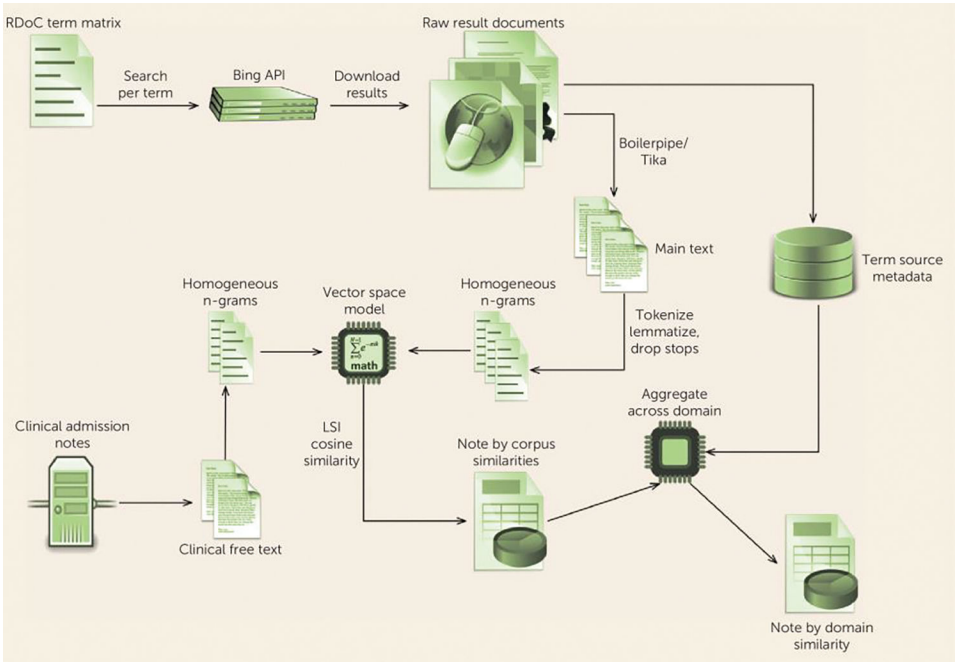

FIGURE 1. Diagram of the General Approach to Deriving a Research Domain Criteria (RDoC) Model and Scoring a Corpus of Narrative Clinical Admission Notes[a]

[a] The method pipeline developed to calculate estimated RDoC domain scores for free-text narrative notes. We defined each RDoC domain as the symptom terms included in the NIMH workgroup concept matrix for that domain. For each term, we conducted automated Internet searches using Bing API to retrieve the 50 most relevant web pages to construct a corpus representative of each RDoC domain (11). The main content text of these results was extracted using Apache Tika, version 1.4 (12), and the Boilerpipe algorithm (13). We applied punkt (14) and Penn Treebank (15) tokenizers and WordNet-based (16) lemmatization as implemented in Natural Language Toolkit, version 2.0.4 (NLTK; 17) to preprocess the domain-defining corpus text into homogeneous lemma before partitioning into mono-, bi-, and trigrams respectful of the tokenized boundaries. After dropping n-grams that included stop words as defined by NLTK, the remaining bag-of-n-grams model was converted into a vector space model (18) the dimensionality of which was reduced (19) (N=200) via latent semantic indexing (LSI) (20) as implemented in Gensim, version 0.8.9 (21). The resulting vector space model is thus rooted in the five RDoC domains, as each document vector arose from a key concept in the workgroup matrix. To score each narrative note, we transformed the text into the domain-defined vector space model. Then we scored the similarity of the resulting clinical document vector to all of the RDoC concept vectors in the model via cosine similarity (22). To reduce the resulting clinical-document-versus-each-concept-document similarity scores, we maintained metadata on the search term, and thus RDoC domain, of origin for each document vector in the domain defined model. Using this metadata, we partitioned the clinical-document-versus-each-concept-document similarity scores for each clinical document by RDoC domain of origin. Finally, the partitioned scores were aggregated into per-domain similarity scores by arithmetic mean, thereby creating five estimated RDoC similarity scores (one per domain) for each clinical document.

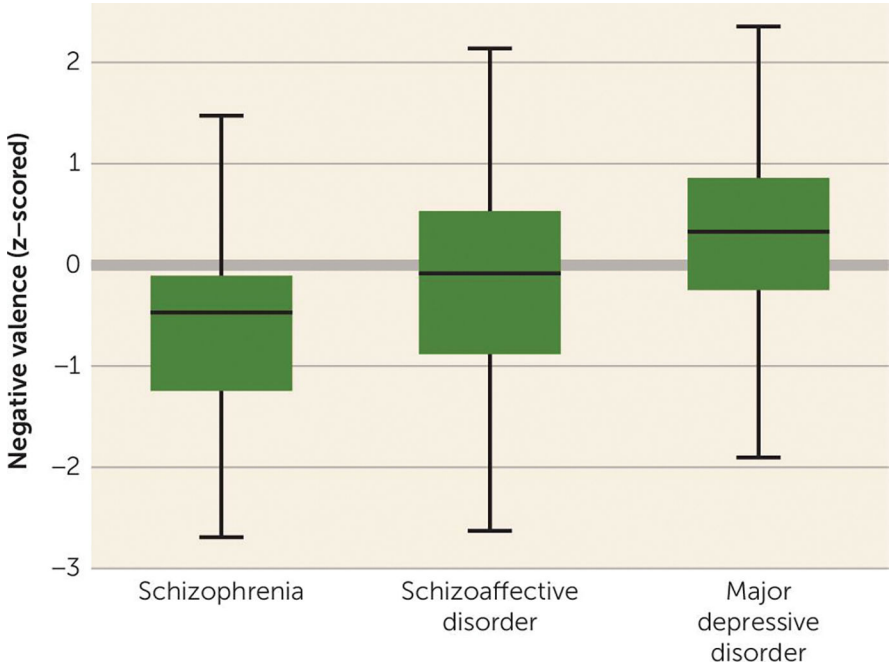

FIGURE 2. Distribution of RDoC Negative Valence Scores Among Individuals in the Cohort With a Diagnosis of Schizophrenia (N=80), Schizoaffective Disorder (N=81), or Major Depressive Disorder (N=327)[a]

[a] Error bars indicate 1.5 interquartile range above and below median; boxes indicate 1st and 3rd quartiles.

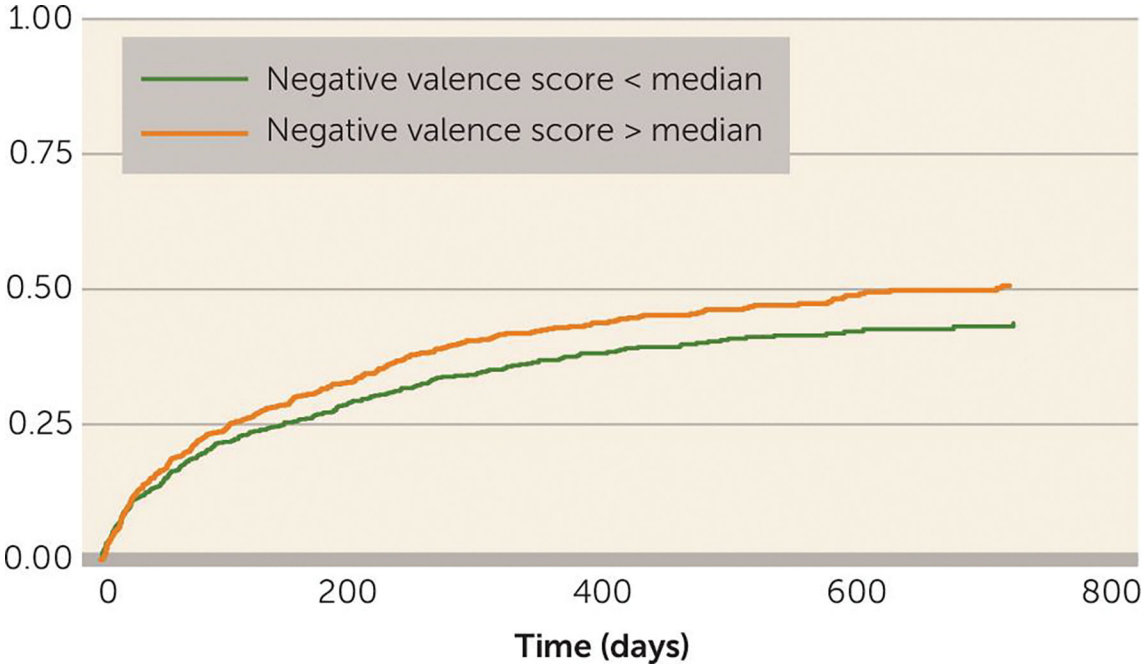

FIGURE 3. Kaplan-Meier Survival Curve for Time to All-Cause Hospital Readmission Based on Median-Split Negative Valence Score, Adjusted for Demographic Features and Other Domain Scores

TABLE 1.

Demographic and Clinical Characteristics of the Psychiatric Admission Cohort[a]

Characteristic		
	N	%
Male	984	49.0
White	1,433	71.3
Public insurance	1,169	58.2
Diagnosis of major depressive disorder	327	16.3
Diagnosis of schizoaffective disorder	81	4.0
Diagnosis of schizophrenia	80	4.0
	Mean	SD
Age (years)	43.8	16.6
Charlson comorbidity index	3.4	4.5
Length of hospital stay (days)[b]	9.1	8.4

[a]The inpatient cohort comprised 2,010 patients for whom there were 2,484 admissions.

[b]Length of hospital stay was available for only 1,862 admissions.

TABLE 2.

Mixed-Effects Regression Examining Association Between Length of Hospital Stay (N=2,315 Admissions) and Individual Estimated Research Domain Criteria (eRDoC) Domain Scores

Variable	Coefficient	SE	z	p	95% CI
eRDoC domains[a]					
Negative valence	−0.88	1.49	−0.59	0.558	−3.81, 2.05
Positive valence	0.21	1.06	0.20	0.843	−1.87, 2.29
Cognitive	3.55	1.06	3.36	0.001	1.48, 5.62
Social	−1.16	1.14	−1.01	0.310	−3.41, 1.08
Arousal	−2.25	1.09	−2.07	0.039	−4.38, −0.12
Demographic and clinical variables					
Age (years)	0.10	0.01	7.13	<0.001	0.07, 0.12
Male	−1.05	0.37	−2.84	0.005	−1.77, −0.32
White	−0.38	0.41	−0.93	0.350	−1.18, 0.42
Public insurance	−0.27	0.34	−0.79	0.428	−0.94, 0.40
Charlson comorbidity index	−0.12	0.05	−2.47	0.013	−0.21, −0.02

[a]With the addition of terms for the eRDoC domains, likelihood-ratio χ^2=32.75, df=5, p,0.001.