Archival Report

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