Electronic health records reveal transdiagnostic clinical features and diverse trajectories of serious mental illness

Juan F. De la Hoz[1], Alejandro Arias[2], Susan K. Service[1], Mauricio Castaño[2], Ana M. Diaz-Zuluaga[1], Janet Song[1], Cristian Gallego[2], Sergio Ruiz-Sánchez[3], Javier I Escobar[4], Alex A. T. Bui[5], Carrie E. Bearden[1], Victor Reus[6], Carlos Lopez-Jaramillo[3], Nelson B. Freimer[1], Loes M. Olde Loohuis[1].

1. Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, USA
2. Department of Mental Health and Human Behavior, University of Caldas, Manizales, Colombia
3. Department of Psychiatry, University of Antioquia, Medellín, Colombia
4. Global Health, Robert Stempel School of Public Health and Social Work, Florida International University, Miami, USA
5. Department of Radiological Sciences, University of California Los Angeles, California, USA
6. Department of Psychiatry and Biobehavioral Sciences, University of California San Francisco, San Francisco, USA

Abstract (max 250)

Objective: Electronic health record (EHR) databases enable scalable investigations of serious mental illness (SMI), including bipolar disorder (BD), severe or recurrent major depressive disorder (MDD), schizophrenia (SCZ), and other chronic psychoses. The authors analyzed structured and unstructured EHR data from a large mental health facility to characterize SMI clinical features and trajectories.

Methods: Diagnostic codes, information from clinical notes, and healthcare use data, were extracted from the EHR database of Clínica San Juan de Dios in Manizales, Colombia for the years 2005-2022, including 22,447 individuals (ages 4-90, 60% female) treated for SMI. The reliability of diagnostic codes was assessed in relation to diagnoses obtained from manual chart review (n=105). A Natural Language Processing (NLP) pipeline was developed to extract features from clinical notes. Diagnostic stability was quantified in patients with ≥ 3 visits (n=12,962). Finally, mixed-effect logistic regression models were used to identify factors associated with diagnostic stability.

Results: Assigned EHR diagnoses showed very good agreement with those obtained from manual chart review (Cohen's kappa 0.78). The NLP algorithm (which demonstrated excellent balance between precision and recall with average F1=0.88) identified high frequencies of suicidality and psychosis, transdiagnostically. Most SMI patients (64%) displayed multiple EHR diagnoses, including switches between primary diagnoses (19%), comorbidities (30%), and combinations of both (15%). Predictors of changes in EHR diagnoses include Delusions in clinical notes (OR=1.50, $p=2e^{-18}$) and a history of previous diagnostic changes (OR=4.02, $p=3e^{-250}$).

Conclusions: Longitudinal EHR databases enable scalable investigation of transdiagnostic clinical features and delineation of granular SMI trajectories through the integration of information from clinical notes and diagnostic codes.

Introduction

Examination of disease trajectories through longitudinal observation of symptoms led to the development of modern classification systems for mental disorders; such data formed the main basis for differentiating key categories of serious mental illness (SMI), such as schizophrenia (SCZ), bipolar disorder (BD), and major depressive disorder (MDD). While these classification systems advocate a parsimonious, longitudinal perspective, current research on SMI relies primarily on cross-sectional assessments, usually of patients with a unique diagnosis, in which the only available trajectory information is supplied by patient recall [1–3]. This lack of detailed longitudinal data may be a factor contributing to heterogeneity within our current SMI categories [4], as evidenced in cross-disorder genetic studies [3] Furthermore, concentrating on individual diagnoses ignores the fact that many features of psychiatric illness (such as suicidality or psychosis) are important transdiagnostically.

Recent studies using longitudinal data collected from participants in national registries [5,6], precision health initiatives [7], and birth cohorts [8], have begun to identify risk factors for specific diagnoses and to describe patterns of variation in these diagnoses over time [5–8]. These resources, which are mainly limited to upper-income countries (UIC), typically contain only sparse data for individual clinical features, including symptoms and behaviors. In contrast, electronic health record (EHR) data which are available in both UIC and in many low- or middle–income countries (LMIC), may contain extensive, descriptions of such clinical features during the periods when individuals actually experience them. EHR databases thus facilitate investigations of features that are important both transdiagnostically and longitudinally, and that may predict clinically important outcomes, such as the onset of psychosis or suicidal behaviors [9,10].

We demonstrate here that the EHR database from a psychiatric hospital located in a middle-income country, enables longitudinal population-scale investigations of SMI diagnoses, individual clinical features, and trajectories [11], through the combination of novel NLP methodologies applied to detailed clinical notes with analyses of structured data. The Clínica San Juan de Dios in Manizales (CSJDM) [12] implemented an EHR in 2005, which provided us both structured and unstructured data from all visits of more than 20,000 SMI patients, from that date until June 2022. The CSJDM provides comprehensive mental healthcare to the one million inhabitants of Caldas, and its EHR captures data concerning SMI covering the entire region[11]. We characterized trajectories of SCZ, BD, and MDD through longitudinal analyses of diagnoses (from assigned diagnostic codes) and of four features (Suicidal Ideation, Suicide Attempt, Delusions, and Hallucinations) described in clinical notes recorded in the EHR, each of which may be important in categorizing SMI and its trajectories, transdiagnostically.

To conduct these analyses, we extracted diagnoses and developed a Natural Language Processing (NLP) pipeline for extraction of transdiagnostic features from the free-text, in Spanish. We first established the reliability and completeness of the EHR and our phenotype extraction pipelines, and showed that features recorded in the notes at individual visits align with ICD-10 diagnostic severity qualifiers at those visits. We then characterized trajectories of each SMI diagnosis, distinguishing between diagnostic *switches* and the accumulation of *comorbidities*, and identifying both global and diagnosis-specific patterns for each. We quantify the probability of patients changing diagnoses across admissions or visits and identify factors contributing to such changes. In particular, we evaluate the utility of transdiagnostic features from narrative notes in delineating SMI trajectories.

Methods

EHR database

To investigate SMI trajectories we extracted from the EHR both structured data (demographic information; duration, type and site of visits [inpatient, outpatient, or emergency department]; diagnostic codes [ICD-10]), and unstructured data, consisting of free-text from clinical notes. These notes include psychiatrists' intake, progress, and discharge notes and nursing notes (from inpatient hospitalizations); psychiatrists' outpatient notes, and psychiatrists' triage notes (from emergency department visits).

Prior to performing the analyses reported here, we removed from the EHR any fields considered Protected Health Information [13], using procedures approved by the institutional review boards from UCLA and CSJDM. We used regular expression matching to strip from the text names and numbers exceeding five digits (potential ID numbers), to further reduce the possibility of including identifying information in our dataset.

For our analyses, we first identified all patients in CSJDM with at least one clinical note (n=77,538) and excluded patients with missing gender information (n=626). We then excluded visits outside the age range of 4-90, without a valid diagnostic code or with primary diagnostic codes outside of the Mental, Behavioral, and Neurodevelopmental Disorders categories (excluded n=20,982 visits from 5,056 patients, Supplementary Figure 1).

ICD-10 codes extraction and cohort definition

Following each visit to the hospital, a patient is assigned a single primary ICD-10 diagnosis by their treating psychiatrist, generating a time-stamped sequence of diagnoses. We extracted this sequence for every patient and selected for analyses patients who had at least one primary diagnosis of SMI,

defined here as BD (F301, F302, F310, F311, F312, F313, F314, F315, F316, F317), Severe/Recurrent MDD (F322, F323, F331, F332, F333, F334), SCZ (F20X), and other chronic psychoses (Delusional Disorder; F22X. Schizoaffective Disorder; F25X). (Supplementary Table 1). In total, this cohort includes 22,447 patients with 157,003 visits (Supplementary Figure 1).

Primary diagnosis classification and reliability estimation

We assessed, in a subsample of the 22,447 patients described above, the reliability of the ICD-10 diagnoses recorded in the EHR in comparison with those made by an expert research clinician (MC) performing a complete manual chart review. To enable a sufficiently precise estimation of the degree of agreement between these two sets of diagnoses, we selected 120 patients for this record review, chosen at random from among participants whom we had previously recruited at CSJDM in an ongoing study of BD, MDD, and SCZ [12]. Of these individuals (40 from each of the three diagnostic groups), we excluded 15 whose most recently recorded ICD-10 diagnoses (F318, F319, F321, F328, or F339) were not among the codes that met our criteria for SMI, as defined above. The clinician review of the remaining 105 records yielded a checklist of symptoms and other clinical features of SMI (see below) and assignment of a *current* primary diagnosis based on DSM-5 criteria [14]. We then, in these 105 patients, evaluated (using Cohen's kappa statistic [15]) the agreement between the diagnosis assigned through this review and the *most recently* recorded ICD-10 SMI code.

NLP algorithm to extract clinical features

To enable the identification of specific clinical features in the clinical notes of each of the 22,447 patients with an SMI primary diagnosis, in the CSJDM database, we developed a Spanish-language NLP algorithm; the procedures used to develop, train, and validate this algorithm are detailed in Supplementary Note 1 (also see Supplementary Tables 2 and 3). For this study, we used the algorithm to identify the presence of four transdiagnostic features that are routinely assessed in clinical encounters: Suicide Attempts, Suicidal Ideation, Delusions, and Hallucinations.

Briefly, two clinicians independently reviewed a randomly selected sample consisting of 3,600 passages of free text (which we term "sentences") from the inpatient notes, progress notes, and outpatient notes of 2,788 unique patients with ICD-10 SMI codes, flagging those sentences in which any of the four features were present. We stopped sentence annotation at this point, as we had identified a sufficient set of positive instances of each feature to obtain accurate estimates of the algorithm's performance.

We evaluated the algorithm's performance using a held-out, gold standard set of 290 sentences, in which each putative feature was also annotated as being affirmed or negated. Metrics used to assess

performance are precision (positive predictive value), recall (sensitivity), and their harmonic mean (F1).

Then, we designated features as "present" for a given patient, if they were identified by the algorithm as having been mentioned affirmatively in a least two notes over the entire course of their EHR. The requirement of two notes was selected as it yielded optimal performance of the NLP algorithm (Supplementary Figure 2). To determine the accuracy with which the algorithm designated the presence or absence of each feature, we compared its output (for the patients in whom we had conducted a manual chart review, as described in the section above), to the checklist of these features compiled by the clinician from the chart review (using the same metrics as for the sentence level evaluation, described above). Finally, as an evaluation of the attributions made by the clinician conducting in the manual review, we conducted an additional set of manual reviews of selected records; two additional clinicians conducted independent reviews of each of the charts with false positive instances (and an equal number of randomly chosen charts with true positives, Supplementary Note 1).

Characterization of extracted clinical features in relation to ICD-10 codes

We evaluated, in the entire study cohort, the correspondence between the clinical features extracted from notes using the NLP algorithm and the current-state severity qualifiers in the ICD-10 diagnoses recorded in the EHR at each individual visit. For these comparisons we used a mixed-effect logistic regression, which accounts for multiple visits per person. Additionally, we include covariates to correct for potential confounding factors such as inpatient status.

Patient-level associations between clinical features and ICD-10 diagnoses

We assessed the relationship, at the individual level, between the presence of clinical features at any timepoint in the EHR, the most recently recorded diagnoses (considering only codes for MDD, BD, and SCZ) and gender. For this analysis, we evaluated all patients with an SMI diagnosis and at least two separately recorded clinical notes (n=20,658 out of 22,447 SMI patients, Supplementary Figure 1). Association tests were performed using logistic regression including both diagnosis and gender while adjusting for the length of patients' records and hospitalization history (Supplementary Note 2). We tested four models, one for each feature. To test for interactions between gender and diagnosis, we expanded the model to include an interaction term.

To evaluate the relationship between clinical features that co-occur in patients (considering their entire longitudinal EHR) we used the same logistic modelling framework, but added, for each feature, the presence, recorded at any point, of the three remaining features, in the same model. To test for

interactions between gender and the second co-occurring feature, we expanded the model to include an interaction term.

Diagnostic switches, comorbidities, and trajectories

We defined two types of diagnostic changes: diagnostic switches and comorbidities. We use the term *diagnostic switches* to refer to changes between two psychiatric diagnoses that cannot, by definition, be held at the same time, specifically, the diagnoses in the ICD-10 F3 and F2 chapters (mood and psychotic disorders, respectively; see Supplementary Note 3 and Supplementary Table 4 for details). By contrast, we use the term *comorbidities* to refer to all other combinations of ICD-10 codes; comorbid diagnoses can accumulate over time, without limit. Using these two definitions, an individual patient's diagnostic trajectory may include both switches and comorbidities.

Diagnostic Stability

We assessed the stability of diagnostic categories over time, considering a diagnosis unstable only if a patient switched to another diagnosis (Supplementary Table 4). For individual SMI diagnoses, we estimated long-term prospective and retrospective stability in those individuals with > 10 visits, (n=12,962, Supplementary Figure 1); we chose this number of visits to represent trajectories of sufficient length for analysis of stability to be meaningful [16,17]. Prospective stability is the probability of a patient's first diagnosis being the same as their last diagnosis, and is analogous to the precision of the initial diagnosis in predicting the final diagnosis. Retrospective stability is the probability of a patient's final diagnosis being the same as their first one and is analogous to recall of the first diagnosis relative to the final. Differences in stability across diagnoses and age groups (before or after age 30) were evaluated using z-tests (Supplementary Note 4).

Factors affecting diagnostic stability

Next, we explored factors contributing to visit-to-visit diagnostic stability. Specifically, we evaluated the effects of patient sex and age, primary diagnosis, inpatient status, previous switch, clinical features, and receiving a Not Otherwise Specified (NOS) code at the previous visit. For these analyses we first used a mixed-effect logistic regression to estimate the probability of switching diagnoses over time (using number of visits as a proxy), accounting for repeated patient observations. We then expanded this model to evaluate the effects of the demographic and clinical factors listed above (Supplementary Note 5). An NOS code indicates diagnostic uncertainty in cases of atypical or confusing patient presentations, or when temporal criteria are not yet met [18], and therefore serves as a positive control; we expect to see

increased diagnostic instability associated with these codes. In a sensitivity analysis, we explored the impact of measuring time by years since the first encounter rather than by visit number.

Finally, to evaluate the possibility that clinical features extracted from the notes at a given visit *anticipate* specific diagnostic changes recorded in future visits, we tested whether psychosis features (Delusions and Hallucinations) predict the application of psychosis current-state qualifiers in ICD-10 diagnoses of BD or MDD (Supplementary Note 6).

Significance thresholds

We applied Bonferroni correction for multiple testing in all our analyses. Model details with corresponding significance thresholds are described in Supplementary Notes 2-6.

Results

Study sample

As of June 2022, the CSJDM EHR included 157,003 visits from 22,447 patients who were assigned an SMI diagnosis at any point from their first visit onwards (Supplementary Figure 1). The demographic and clinical characteristics of this sample are described in Supplementary Table 5.

Reliability of diagnoses and clinical features extracted from EHR compared to those identified through manual chart review by expert clinicians

For the 105 randomly selected patients in whom we conducted a complete manual chart review, the diagnoses extracted from their EHR (most recent assigned ICD-10 code) demonstrated agreement with their diagnoses from manual review of their entire EHR at a kappa level typically considered "very good" to "excellent" for such comparisons [19,20]. The kappa estimates for specific diagnoses considering both inpatient visits and outpatient visits (Supplementary Table 6) were: 0.74 (95% CI: 0.60-0.89) for MDD, 0.74 (95% CI: 0.60-0.87) for BD, 0.90 (95% CI: 0.81-0.99) for SCZ; overall kappa=0.78 (95% CI: 0.69-0.88). Positive predictive values (PPVs) were 0.84 for MDD, 0.80 for BD, and 0.92 for SCZ. Estimates for kappas and PPVs were similarly high when considering inpatient visits only (Supplementary Table 6). These levels of agreement are also similar to those from previous studies comparing diagnoses from ICD codes recorded in the EHR with diagnoses from manual chart reviews [21].

In training our NLP algorithm for extracting clinical features from the EHR notes, the kappas indicated "good" to "excellent" agreement between two independent annotators for all four features (Supplementary Table 7). Then, application of the algorithm to extract features from the gold standard set

of sentences, demonstrated that it performed with a high rate of precision (range: 0.88-1.0) and recall (range: 0.62-1.0), resulting in a satisfactory F1 for all features (Suicide Attempt: 0.82, Suicidal Ideation: 0.73, Delusions: 1.0, Hallucinations: 0.95), (Supplementary Table 8A, see Supplementary Note 7 for a description of errors). We evaluated different thresholds for the number of affirmative mentions of a feature in the output of the NLP algorithm that we would require to consider that feature "present", for a given patient, over the lifetime of their EHR (Supplementary Figure 2); requiring two such mentions provided an optimal balance (as measured by F1) between precision and recall. At this threshold the algorithm output and the designation of features from manual chart review were highly concordant for all four features; Suicide Attempt (92/104, F1 = 0.68), Suicidal Ideation (89/104, F1 = 0.79), Delusions (84/104, F1 = 0.82), and Hallucinations (87/104, F1 = 0.84), Supplementary Table 8B. Further investigation suggested that the above comparison actually underestimated the performance of the algorithm (Supplementary Note 1).

Comparison between ICD-10 diagnoses of MDD and BD assigned at each visit with clinical features identified from the notes from the same visit

The ICD-10 codes for MDD and BD include qualifiers indicating the severity of episodes (unipolar depressive, bipolar depressive, and manic) and the presence or absence of psychotic features at a given visit; this information provides the opportunity to evaluate the relationship between these qualifiers and the clinical features extracted from the notes using the NLP algorithm. As would be expected, we observe strong positive associations for all four of the extracted clinical features with both episode severity and presence of psychosis, as recorded in the ICD-10 codes (Supplementary Note 6, Supplementary Table 9).

In contrast to the ICD-10 codes, which simply report the presence or absence of psychotic symptoms at a given visit, the application of the NLP algorithm to the clinical notes for these visits provides rich information on such symptoms. The notes reveal that, in depressive episodes (both unipolar and bipolar) Delusions and Hallucinations are observed at relatively similar frequencies, while in manic episodes Delusions represent by far the predominant psychotic feature (Figure 1). This finding is consistent with observations made in a manual review of 1,715 case notes from a North American tertiary care hospital [22], while a systematic review article of types of psychotic symptoms in bipolar depression and mania reported less consistent pattern across multiple, smaller studies [23].

The notes contain corroborating examples for most of the instances in which an ICD-10 diagnosis of depression or mania with psychotic features was assigned; for visits at which such diagnoses were recorded, the algorithm identified either Hallucinations or Delusions at a frequency ranging from 72%

(unipolar depression) to 84% (mania, see Figure 1A). Examples of such psychotic features were also found, however, in a substantial proportion of the notes of visits for which the recorded ICD-10 diagnosis indicated an absence of psychotic features (ranging from 21% [unipolar depression] to 56% [mania]), Figure 1B). It is unclear what accounts for this apparent discrepancy, but one contributing factor could be differences between the clinicians recording ICD-10 diagnoses and the NLP algorithm in how they consider psychosis. Such differences could at least partially explain the high frequency of Delusions in the notes from visits for which the diagnosis of "mania without psychotic features" was assigned (47%). Grandiose delusional beliefs comprise more than 30% of the Delusions for such visits identified by the NLP algorithm (Supplementary Figure 3, but, in practice, the point at which grandiosity (a cardinal feature of mania) reaches psychotic proportions may be difficult to define [24]. Additionally, persistence of delusionary beliefs may have been a consideration for clinicians recording ICD-10 diagnoses, while in some instances the information extracted from the notes may reflect transient beliefs that resolved relatively quickly.

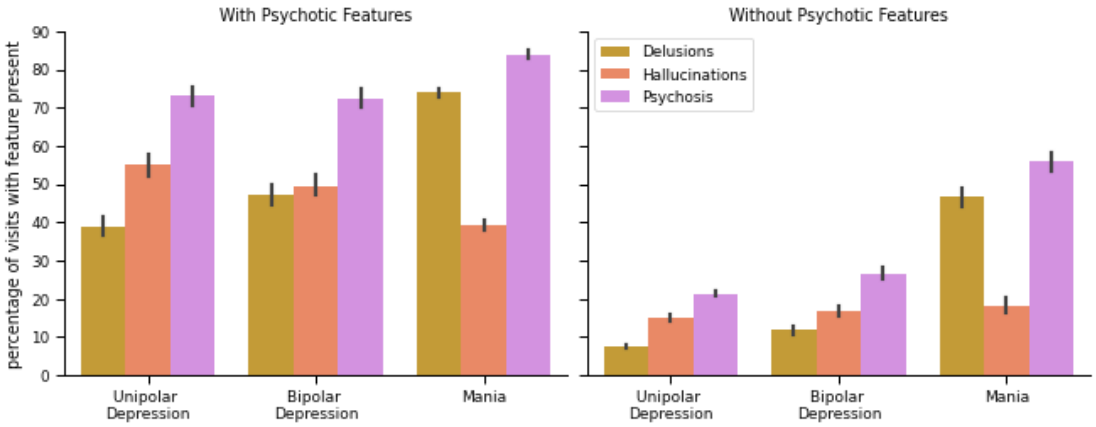

Figure 1. Psychotic features extracted from the clinical notes of visits for severe unipolar and bipolar depression and mania. The ICD-10 codes for these disorders (F32, F33, F31) include qualifiers for the clinician to specify the presence or absence of psychosis for each visit. (1A) The percentage of visits assigned a diagnosis of "with psychotic features" for which the NLP algorithm identified the features Delusions and Hallucinations, considered together ("Psychosis") and separately. (1B) The percentage of visits assigned a diagnosis of "without psychotic features" for which the NLP algorithm identified the features Delusions and Hallucinations, considered together ("Psychosis") and separately. Error bars indicate 95% confidence intervals obtained through bootstrapping, n=5,961 patients and 13,928 visits.

Transdiagnostic characterization of features extracted from EHR notes

The above comparisons indicated that the NLP algorithm identifies the presence of psychosis and suicidality clinical features in the EHR database with high sensitivity. Suicide Attempt, Suicidal Ideation, Delusions, and Hallucinations each occur in all of the SMI diagnoses at a frequency of >5%, stratified by gender, demonstrating their transdiagnostic quality (Figure 2A). Several patterns of these frequencies are, however, noteworthy. In contrast to most reports in the literature [25,26], Suicidal Ideation in the CSJDM database is less frequently observed in females compared to males, after correcting for diagnoses, inpatient history and number of visits (OR=0.84, $p=8.42e^{-7}$, Supplementary Tables 10 and 11). This difference is driven largely by a lower rate of Suicidal Ideation in females with MDD specifically (34% versus 45% in males, interaction OR=0.65, $p=4.2e^{-9}$,). A similar reduced frequency of psychotic features was observed in females (OR=0.67, $p=1.99e^{-22}$ Delusions; and OR=0.88, $p=5.9e^{-4}$ Hallucinations,), while rates of Suicide Attempt are similar in both genders. The four features subdivide the patient population according to the combination of comorbidities (Figure 2B). Aside from the expected co-occurence of suicide-related features and psychotic features, we found, unexpectedly, that the mention of Delusions in the notes decreases the likelihood of notes mentioning Suicidal Ideation or Suicide Attempts in the same patient and vice versa (OR between 0.59-0.62, $p < 1.78e^{-17}$, accounting for gender, diagnosis, inpatient history and number of visits [Supplementary Table 11); the reverse is true for Hallucinations (OR between 1.29-2.05, $p < 2.89e^{-7}$).

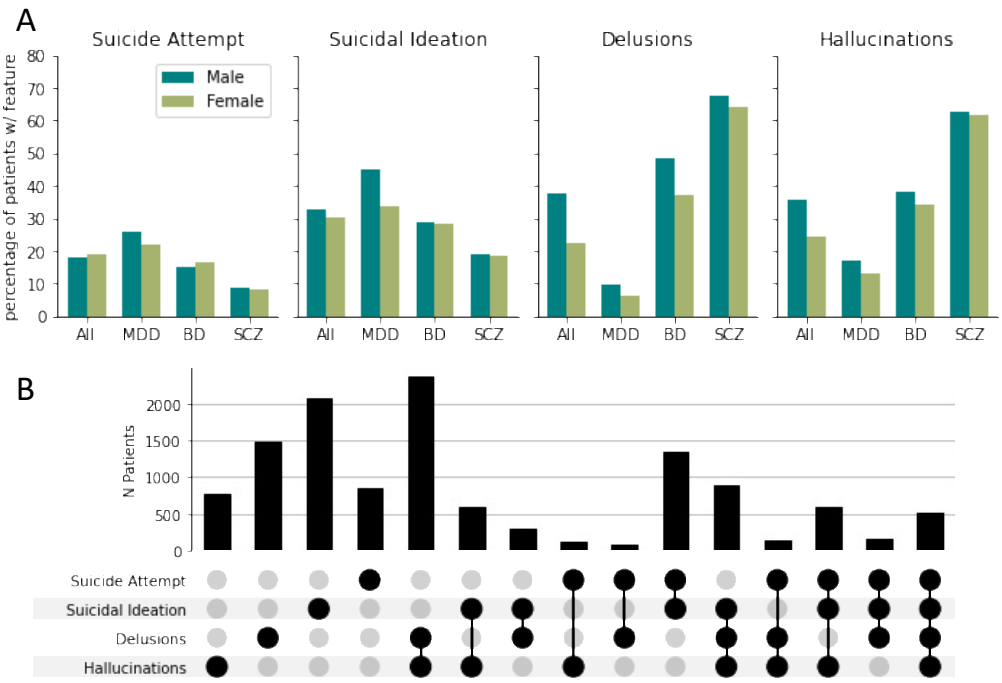

Figure 2. Transdiagnostic characterization and co-occurrence of clinical features extracted from EHR notes

A) Proportion of patients with each of the four features stratified by primary diagnosis. B) Number of patients with co-occurrence of 2, 3, or 4 clinical features. All data in these plots are limited to patients with at least two EHR notes.

Diverse diagnostic trajectories in SMI patients

We evaluated diagnostic trajectories among all SMI patients with at least three visits (n=12,962, Supplementary Figure 4). The majority (64%, Figure 3A) had multiple diagnoses recorded in their EHR, broken down as follows: 30% displayed comorbidities (orange bars; Supplementary Table 12, 19% displayed diagnostic switches (teal bars), and 15% displayed both switches and comorbidities (purple bars).

While some pairs of diagnoses in the trajectories are common, for example, the switch from MDD to BD (observed in 24% of current BD patients, figure 3B) and the comorbidity between MDD and Other Anxiety Disorders (observed in 28% of current MDD patients), the majority of patients (58%) follow rare trajectories (occurring in fewer than 1% of patients). Altogether, we observed 3,149 unique trajectories.

We estimated prospective and retrospective stability for each diagnosis, evaluating trajectories with 10 or more visits (n=5,016). Prospective stability was lower for MDD compared to BD or SCZ (56% vs. 88% and 83%, respectively; 2-df chi-square=383; $p=5e^{-84}$). Retrospective stability, by contrast, while lower than prospective stability for all diagnoses, was highest in MDD (53% vs. 48% and 40% in BD and SCZ respectively; 2-df chi-square=34.5; $p=3e^{-8}$). (Supplementary Note 4 and Supplementary Table 13).

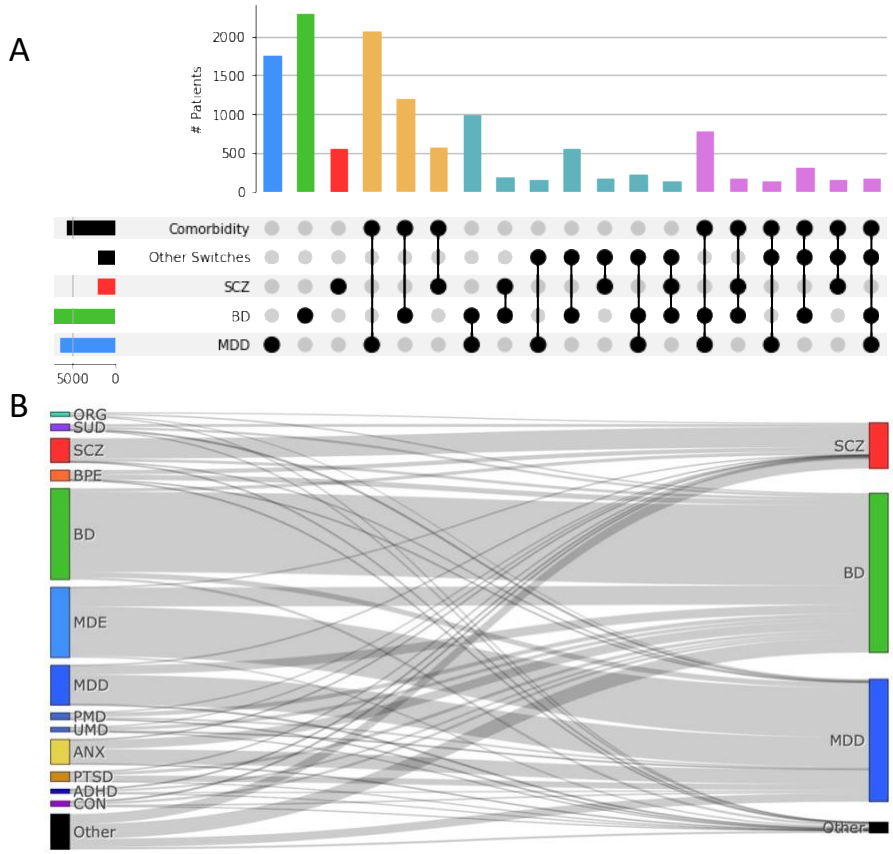

Figure 3: Disease trajectories of SMI in patients with at least three visits. A) UpSet plot presenting diagnostic switches (between SMI categories) and comorbidities (SMI and non-SMI categories). Patients with a single SMI diagnosis (blue, green, red, total n=4,620); a single SMI diagnosis and other comorbidities (orange n=3,955); multiple SMI diagnoses and no other comorbidities (teal n=2,468); multiple SMI diagnoses and other comorbidities (purple, n=1,919). Bars with n<100 are not shown. B) Sankey diagram of ICD-10 code trajectories. Left nodes represent the diagnosis given at the initial visit and right nodes represent the most recent SMI code. (Diagnostic switches within SMI are shown in Supplementary Figure 4). ORG: Other mental disorders due to brain damage and dysfunction and to physical disease (F06), SUD: Mental and behavioral disorders due to multiple drug use and use of other psychoactive substances (F19), BPE: Acute and transient psychotic disorders (F23), MDE: Major Depressive Episode (F32), PMD: Persistent mood disorders (F34), UMD: Unspecified mood disorder (F39), ANX: Other anxiety disorders (F41), PTSD: Reaction to severe stress, and adjustment disorders (F43), ADHD: Hyperkinetic disorders (F90), CON: Conduct disorders (F91)

Clinical features, time, and other factors affecting diagnostic stability and specific trajectories

We identified multiple factors that influenced diagnostic stability. Diagnostic switching was most frequent during the early stages of treatment. While 11.3% of the patients changed diagnosis on their second visit, this percentage decreased over the patient's course of illness (Figure 4A; log10(k) OR=0.56, p-value $5e^{-66}$) and stabilized at around 4% after the tenth visit. Additional predictors of future diagnostic

instability included the following observations at the current visit: a diagnostic switch from the previous visit (Figure 4B; OR=4.02, p-value $3e^{-250}$), an inpatient visit (OR=1.7, p-value $5e^{-35}$), an NOS diagnosis (OR=1.61, p-value $2e^{-47}$), and the presence of the clinical features Delusions or Hallucinations (OR=1.50 and 1.17, p-values $2e^{-18}$ and $3e^{-4}$, respectively). Predictors of future diagnostic stability included: diagnoses of SCZ or BD compared to MDD (ORs=0.31 and 0.32; p-values $<3e^{-70}$), male gender (OR=0.71, p-value $2e^{-16}$), and increasing age (OR per decade=0.96, p-value $8e^{-4}$). Sensitivity analyses confirmed these findings; the same pattern was observed when modeling switching by time rather than visit number (Supplementary Figure 5).

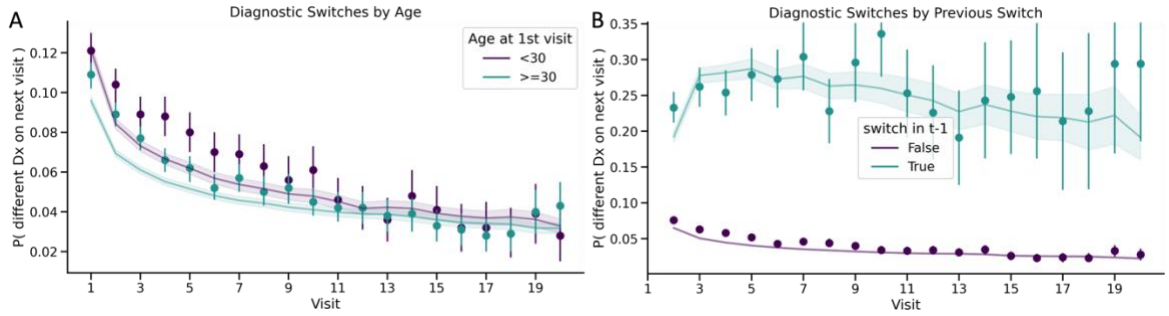

Figure 4: Diagnostic stability over time. At each visit k, the proportion of patients that will switch primary diagnosis code on their next visit k+1. A) Stratified by age groups: age at 1st visit before and after 30 years. B) Stratified by having previously switched diagnoses (from visit k-1). n=12,962 patients (Supplementary Figure 1).

Discussion

By analyzing EHR data spanning 17 years and encompassing over 20,000 patients from a single large mental health facility, we characterize SMI, its relation to transdiagnostic clinical features, and its longitudinal trajectories. We show that both the diagnostic codes recorded in the EHR and our custom, rule-based NLP pipeline for extracting clinical features from narrative notes, reliably reflect the diagnostic impressions of expert clinicians conducting manual chart reviews. By applying this pipeline to our data we were able to perform more granular analyses of four key clinical features of SMI – Suicide Attempts, Suicidal Ideation, Delusions, and Hallucinations – than has been possible in most previous EHR studies of SMI which have relied only on the information in the diagnostic codes. Additionally, analysis of the NLP-algorithm output in relation to the recorded diagnostic codes reveals the high degree of association between the information contained in these two types of data. The four clinical features each occur at a high frequency transdiagnostically, and they co-occur in both expected and unexpected patterns. Finally, we use the information on clinical features in the notes together with the ICD-10

diagnostic codes to characterize SMI trajectories, including prediction of future diagnostic changes, differentiation of diagnostic switches from the accumulation of comorbidities, and factors contributing to the stability of diagnoses.

Our NLP pipeline overcomes the performance and usability issues of "off-the-shelf" NLP pipelines [27,28] and is, to our knowledge, the first of its kind for extracting information from Spanish language psychiatric notes. The algorithm identified the clinical features in the notes from multiple individuals where the initial chart review failed to do so, and its results were mostly confirmed by a further round of manual reviews. These observations suggest, therefore, that this automated approach to retrieving clinical data is more accurate than manual review, as well as being vastly more scalable.

In considering the data on psychotic features that we obtained from the clinical notes in relation to what we gleaned from the ICD-10 codes, we illustrate the potential utility of the NLP pipeline for analyzing a vast array of information that is often unavailable in large-scale studies of SMI, for which the phenotypes may be based on brief interviews or self-report scales (e.g., [29]). Notably, several previous studies have reported conflicting results with respect to the relative frequencies of Delusions and Hallucinations in psychotic mania compared to psychotic depression, which may reflect their differing designs and mostly small sample sizes of the studies that they have evaluated [23]. In contrast, by applying a uniform methodology to analyze information extracted from the notes of nearly 6,000 patients, in almost 14,000 clinical encounters, we were able to show that Delusions are far more frequent than Hallucinations in mania, while Hallucinations have the same or even greater frequency than Delusions in both unipolar and bipolar depression. This observation only partially reflects grandiose delusional beliefs characteristic of mania. Further development of the NLP algorithm will enable even finer-grained characterization of psychosis in these disorders, e.g., specifying whether psychotic symptoms in the notes are congruent or incongruent with the predominant mood states of an episode [23].

Our comparison of the information from the notes and from the diagnostic codes also highlights the complementarity of these data sources, as, in most instances, the former provide specific examples to corroborate the assignment of the "with psychotic features" codes. On the other hand, the apparent discrepancies between the notes and the codes – the identification of psychotic features in the free text at visits where such features were not recorded in the diagnoses – provide a reminder that the EHR is primarily a clinical record and the contents of its various components reflect their differing purposes. The assigned diagnostic codes represent a clinician's overall impression of a patient's predominant clinical state at a given visit, rather than a comprehensive representation of all of the clinical information obtained at that visit.Even though such information may not be incorporated within the formal diagnosis, it is, however, of great value for addressing a number of important questions,.

Widespread recognition of the inadequacies of diagnosis-based taxonomies for classifying mental illness has generated growing interest in investigation of transdiagnostic features of SMI [30]. Evidence has accumulated indicating a high degree of shared genetic risk across SMI diagnoses, but assembling adequately-scaled datasets that are suitable for genetic analysis of systematically assessed transdiagnostic phenotypes has proven challenging. Our finding that all four of the clinical features that we extracted from the EHRs are present in substantial numbers in each of the SMI diagnostic categories suggests that extending this approach to additional features, and to the EHR databases of additional facilities, could enable well-powered genetic association studies of a number of phenotypes that are relevant to SMI. Further transdiagnostic explorations will also be important to follow up unexpected findings from our current analyses for which we have no obvious explanation, including the lower rates [lifetime EHR] of Suicidal Ideation, Delusions, and Hallucinations identified in females compared to males, and the contrasting patterns of association between the suicidal features and the psychotic features (negative for Delusions, and positive for Hallucinations).

Our findings regarding longitudinal diagnostic trajectories of SMI rest on an EHR database that is, to our knowledge, unique. Because the CSJDM provides comprehensive care to all individuals living with SMI in a geographically defined catchment area, the EHR provides an essentially continuous record of treatment encounters, over this period in the more than 20,000 individuals included in our analyses. These data enabled us to model simultaneously the dynamics of switching between incompatible SMI diagnoses and the accumulation of comorbidities; by doing so, we found that approximately half of the SMI population had one or more psychiatric comorbidities, and over one-third have switched diagnosis at least once. The most frequent comorbities are between anxiety disorders and MDD, while the most frequent diagnostic switch is between MDD and BD. The combination of comorbidities and diagnostic switches describes a broad variety of disease trajectories; a major challenge for future studies will be to determine which of these patterns are most meaningful, either from the standpoint of our understanding of disease causation or in terms of clinical utility. Additionally, as have others we found that diagnostic instability is characteristic mainly of the early stages of SMI [5,31].

Most previous studies of SMI trajectories at a population level have utilized national registries available mainly in a few Northern European countries [5,6] In particular, studies of diagnostic progression in patients with an index psychiatric diagnosis have observed similar degrees of instability of initial SMI diagnoses as we report here [6]. While the concordance of our results with those of these registry studies provides an additional validation for our approach, our integration of features from clinical notes together with diagnostic codes and their qualifiers, has enabled us to delineate trajectories at a level of granularity not available in registry data, and to identify patterns that could have important clinical implications. As

an example, we observed that mentions of psychosis in the notes from a clinical visit significantly preceded diagnostic switches at future visits. More specifically, we noted that examples of Delusions in the notes from a clinical visit at which the ICD-10 qualifier "without psychotic features" was assigned, anticipated the application of the qualifier "with psychotic features" at the subsequent visit. This observation will stimulate further research with the aim of determining how such information could be used for clinical prediction.

Our finding that diagnostic switches increase the likelihood of future switches is driven in large part by a small number of patients who rapidly accumulate diagnoses. While we have not yet identified any specific features that characterize these patients, we hypothesize that they constitute a group for whom research classifications that assign all case-participants a single primary lifetime diagnosis are particularly imprecise descriptions of their phenotype. This group is difficult to recognize in cross-sectional studies, and we hypothesize that they may constitute a source of variance in the many large-scale psychiatric genetics investigations that rely on such classification systems. By contrast, utilizing longitudinal EHR data to selecti samples for genetic studies may not only facilitate the early identification of individuals with extreme diagnostic instability and reduce heterogeneity of research datasets, but alsoenable discovery of distinct features that characterize this group.

Limitations of the study are its focus on a single mental health care facility, the CSJDM, and the early stage of development of the NLP algorithm that we use here to automate the extraction of clinical information from its EHR. As we have already noted, the CSJDM was ideal for implementing our approach due to the extensiveness of its EHR database and given the fact that it continues to provide most of the care for SMI to the ~ 1,000,000 inhabitants of the Department of Caldas. We are now extending the approaches described here, including further testing of the NLP algorithm, to enable longitudinal studies of SMI in other facilities, including the Hospital Mental de Antioquia, one of the largest psychiatric hospitals in Colombia. These implementations will enable estimations of factors such as institutional biases in reporting styles.

Further development of the NLP algorithm to incorporate a larger number of symptoms and behaviors will enable us to place greater confidence in the validity of the assigned ICD-10 diagnoses, e.g., for analyses of trajectories or for future genetic association studies. For example, our current definition of diagnostic switches include individuals who have transitioned from a diagnosis of BD to one of MDD. Such a transition implies either that the clinician assigning the former diagnosis erroneously included an episode of mania or hypomania, or that the clinician assigning the latter diagnosis erroneously overlooked such an episode; our aim is to use the algorithm to help resolve such uncertainties. Additionally, while we currently only differentiate between affirmations and negations of particular terms, in future work we

will incorporate a larger range of contexts that are relevant for delineating trajectories, e.g., distinguishing between symptoms that are improving and those that are worsening over the course of a hospitalization.

Our results support the notion that research classifications that incorporate past and future trajectory data will likely be less heterogeneous and more realistic than current systems that assign patients a single 'lifetime' diagnosis. Evidence from prior studies suggests that distinctive genetic risk profiles may partially underlie trajectory features, such as polarity at onset in BD [32] or conversion from non-psychotic to psychotic illness [33–35]. Efforts to replicate and extend such findings, however, have been limited by variation in ascertainment strategies, reliance on patient recall [32], and small sample sizes [33,35]. Future research should evaluate the relationship between genetic risk and diagnostic or clinical stability, with the aim of establishing more genetically homogeneous subgroups. As analyzing datasets with thousands of uncommon trajectories will be impractical, developing improved methods for reducing dimensionality by clustering patients with similar trajectories should be an important focus of future work[36]. EHR databases usually contain information on interventions, such as pharmacological treatments, that likely influence disease trajectories. Modeling this impact will be an important direction of future research.

Acknowledgments

Research reported here was supported by R01MH123157 (to LMOL, CLJ, and NBF), R01 MH113078 (to CEB, CLJ, and NBF), R00 MH116115 (to LMOL), T32 MH073526 (to JFDLH) and the Fulbright Commission in Colombia under the Fulbright-Colciencias grant (to JFDLH). The content is solely the responsibility of the authors and does not represent the official views of the Fulbright Program or the National Institutes of Health.

References

1. Kendler, K. S. The nature of psychiatric disorders. *World Psychiatry* **15**, 5–12 (2016).
2. Hyman, S. E. The diagnosis of mental disorders: The problem of reification. *Annual Review of Clinical Psychology* vol. 6 155–179 Preprint at https://doi.org/10.1146/annurev.clinpsy.3.022806.091532 (2010).
3. Anttila, V. *et al.* Analysis of shared heritability in common disorders of the brain. *Science (1979)* **360**, (2018).
4. Regier, D. A. *et al.* DSM-5 field trials in the United States and Canada, part II: Test-retest reliability of selected categorical diagnoses. *American Journal of Psychiatry* **170**, 59–70 (2013).
5. Plana-Ripoll, O. *et al.* Exploring Comorbidity Within Mental Disorders among a Danish National Population. *JAMA Psychiatry* **76**, 259–270 (2019).

6. Høj Jørgensen, T. S., Osler, M., Jorgensen, M. B. & Jorgensen, A. Mapping diagnostic trajectories from the first hospital diagnosis of a psychiatric disorder: a Danish nationwide cohort study using sequence analysis. *Lancet Psychiatry* **10**, 12–20 (2023).
7. Barr, P. B., Bigdeli, T. B. & Meyers, J. L. Prevalence, Comorbidity, and Sociodemographic Correlates of Psychiatric Disorders Reported in the All of Us Research Program. *JAMA Psychiatry* **79**, 622–628 (2022).
8. Caspi, A. *et al.* Longitudinal Assessment of Mental Health Disorders and Comorbidities Across 4 Decades Among Participants in the Dunedin Birth Cohort Study. *JAMA Netw Open* **3**, e203221 (2020).
9. Barak-Corren, Y. *et al.* Predicting suicidal behavior from longitudinal electronic health records. *American Journal of Psychiatry* **174**, 154–162 (2017).
10. Raket, L. L. *et al.* Dynamic ElecTronic hEalth reCord deTection (DETECT) of individuals at risk of a first episode of psychosis: a case-control development and validation study. *Lancet Digit Health* **2**, e229–e239 (2020).
11. Song, J. *et al.* Geospatial analysis reveals distinct hotspots of severe mental illness. *medRxiv* (2022) doi:10.1101/2022.03.23.22272776.
12. Service, S. K. *et al. Distinct and shared contributions of diagnosis and symptom domains to cognitive performance in severe mental illness in the Paisa population: a case-control study*. *Articles Lancet Psychiatry* vol. 7 www.thelancet.com/psychiatry (2020).
13. Office for Civil Rights. *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*. (2012).
14. American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5*. vol. 5 (American psychiatric association Washington, DC, 2013).
15. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas* **20**, 37–46 (1960).
16. Schwartz, J. E. *et al.* Congruence of Diagnoses 2 Years After a First-Admission Diagnosis of Psychosis. *Arch Gen Psychiatry* **57**, 593–600 (2000).
17. Baca-Garcia, E. *et al.* Diagnostic stability of psychiatric disorders in clinical practice. *British Journal of Psychiatry* **190**, 210–216 (2007).
18. First, M. B. *et al.* Do mental health professionals use diagnostic classifications the way we think they do? A global survey. *World Psychiatry* **17**, 187–195 (2018).
19. Clarke, D. E. *et al.* DSM-5 Field Trials in the United States and Canada, Part I: Study Design, Sampling Strategy, Implementation, and Analytic Approaches. *American Journal of Psychiatry* **170**, 43–58 (2013).
20. Kraemer, H. C., Kupfer, D. J., Clarke, D. E., Narrow, W. E. & Regier, D. DSM-5: How reliable is reliable enough? *American Journal of Psychiatry* vol. 169 13–15 Preprint at https://doi.org/10.1176/appi.ajp.2011.11010050 (2012).
21. Davis, K. A. S., Sudlow, C. L. M. & Hotopf, M. Can mental health diagnoses in administrative data be used for research? A systematic review of the accuracy of routinely collected diagnoses. *BMC Psychiatry* **16**, (2016).
22. Black, D. W. & Nasrallah, A. Hallucinations and delusions in 1,715 patients with unipolar and bipolar affective disorders. *Psychopathology* **22**, 28–34 (1989).
23. Chakrabarti, S. & Singh, N. Psychotic symptoms in bipolar disorder and their impact on the illness: A systematic review. *World J Psychiatry* **12**, 1204–1232 (2022).

24. Canuso, C. M., Bossie, C. A., Zhu, Y., Youssef, E. & Dunner, D. L. Psychotic symptoms in patients with bipolar mania. *J Affect Disord* **111**, 164–9 (2008).
25. Canetto, S. S. & Sakinofsky, I. The gender paradox in suicide. *Suicide Life Threat Behav* **28**, 1–23 (1998).
26. Murphy, G. E. Why women are less likely than men to commit suicide. *Compr Psychiatry* **39**, 165–75 (1998).
27. Akhtyamova, L. Named Entity Recognition in Spanish Biomedical Literature: Short Review and Bert Model. in *Conference of Open Innovation Association, FRUCT* vols 2020-April 3–9 (IEEE Computer Society, 2020).
28. Cotik Viviana and Rodríguez, H. and V. J. Spanish Named Entity Recognition in the Biomedical Domain. in *Information Management and Big Data* (ed. Lossio-Ventura Juan Antonio and Muñante, D. and A.-S. H.) 233–248 (Springer International Publishing, 2019).
29. Cai, N. *et al.* Minimal phenotyping yields genome-wide association signals of low specificity for major depression. *Nat Genet* **52**, 437–447 (2020).
30. Insel, T. *et al.* Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *Am J Psychiatry* **167**, 748–51 (2010).
31. Bromet, E. J. *et al.* Diagnostic Shifts during the decade Following First Admission for Psychosis. *American Journal of Psychiatry* **168**, 1186–1194 (2011).
32. Kalman, J. L. *et al.* Characterisation of age and polarity at onset in bipolar disorder. *British Journal of Psychiatry* **219**, 659–669 (2021).
33. Perkins, D. O. *et al.* Polygenic risk score contribution to psychosis prediction in a target population of persons at clinical high risk. *American Journal of Psychiatry* **177**, 155–163 (2020).
34. Musliner, K. L. *et al.* Polygenic risk and progression to bipolar or psychotic disorders among individuals diagnosed with unipolar depression in early life. *American Journal of Psychiatry* **177**, 936–943 (2020).
35. Jonas, K. G. *et al.* Schizophrenia polygenic risk score and 20-year course of illness in psychotic disorders. *Transl Psychiatry* **9**, (2019).
36. Krebs, M. D. *et al.* Associations between patterns in comorbid diagnostic trajectories of individuals with schizophrenia and etiological factors. *Nat Commun* **12**, (2021).