Analysis of diagnosis instability in electronic health records reveals diverse disease trajectories of severe 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 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. Department of Psychiatry, Herbert Wertheim College of Medicine, Florida International University, Miami, USA
5. Department of Radiological Sciences, University of California Los Angeles, California, USA
6. Department of Psychiatry, University of California San Francisco, San Francisco, USA

Abstract

Investigations using Electronic Health Records (EHR) databases could enable accurate delineation of psychiatric disease trajectories at an unprecedented scale. Using EHR from a single institution (Clínica San Juan de Dios in Manizales, Colombia), we characterize diagnostic trajectories of >22,000 (ages 4-90, 60% female) individuals treated for severe mental illness (SMI), including schizophrenia (SCZ), bipolar disorder (BD), and severe or recurrent major depressive disorder (MDD).

We extracted diagnostic codes, clinical notes, and healthcare use data collected since 2005. Using a subsample of 105 SMI patients, we assessed diagnostic reliability, comparing EHR to clinical chart review. EHR diagnostic codes showed very good agreement with chart review diagnoses (Cohen's kappa 0.78). Using 3,600 annotated sentences from 2,788 patients, we developed a pipeline for extracting clinical features from the electronic text, which showed high agreement with gold-standard annotations (average F1 0.88). Factors associated with diagnostic instability, defined as changes in diagnosis between successive visits, were identified using mixed-effect logistic regression models.

Of SMI patients with >3 visits (n=12,962), 64% had multiple EHR diagnoses; diagnostic switches (19%), comorbidities (30%), and both (15%). While some diagnostic switches are common, such as the switch from MDD to BD (observed in 22% of BD patients), trajectories are highly heterogeneous, with rare trajectories (occurring in <1% of patients) making up the majority (58% of all patients). Predictors of diagnostic instability include time since initial visit (OR 0.56 by visit number, p-value $2e^{-66}$), previous diagnostic change (OR= 4.02, p-value $3e^{-250}$) and NLP-derived descriptions of delusions (OR 1.50, p-values $2e^{-18}$).

Our results underline the importance of considering longitudinal rather than cross-sectional diagnoses in psychiatric research and show how high-quality EHR data can contribute to global efforts to understand disease trajectories.

Introduction

The delineation of disease trajectories is an important goal of precision health. Identification of individual or group differences in the course of illness may be a powerful means to reduce heterogeneous classifications based on cross-sectional diagnostic assessments into more homogeneous subtypes[1]; reaching this goal will increase our understanding of pathobiology and improve clinical care. Such an advance is particularly important in psychiatry, in which self-report information of discrete symptoms underlies the heterogeneous categories for classifying SMI in modern nosologies[2–4]. In current psychiatric research, such clinical information is mostly collected cross-sectionally. This situation is ironic, as classification systems for mental disorders have their roots in longitudinal observations of trajectory variations in SMI made by pioneer psychiatrists of the late 19th and early 20th centuries[5]. Furthermore, efforts to better characterize and understand such variations have long been a focus of psychiatric research.

Two aspects of SMI trajectories have attracted increased attention in recent years: common diagnostic switches, such as the conversion from MDD to BD [6–8], and common comorbidities, such as that seen between depressive and anxiety disorders[9]. Among the factors stimulating this attention is evidence from studies suggesting that distinctive genetic risk profiles may partially underlie trajectory features such as polarity at onset in BD [10] or conversion from non-psychotic to psychotic illness [11–13]. Efforts to replicate and extend such findings, however, have thus far been limited by variations in ascertainment strategies, reliance on patient recall [10], and small sample sizes [11–13].

Recent studies of SMI trajectories have largely relied on resources available primarily in upper-income countries (UICs), including national registries [14], biobanks [15], and longitudinal birth cohorts [16]. We have recently highlighted the value of using the EHR from an institution located in a low- to middle-income country (LMIC) for population-level investigations of SMI [17]. EHR have the advantage that they are collected passively, and, unlike registries, contain detailed descriptions of symptomatology, providing unprecedented opportunities to delineate SMI trajectories, transdiagnostically.

Here, we aim to characterize SMI trajectories in a dataset with longitudinal and granular information on both diagnoses and symptoms. To do this, we leverage 17 years of EHR data from a mental health provider in the country of Colombia, the Clínica San Juan de Dios in Manizales (CSJDM), serving an entire region with about one million people (department of Caldas)[17]. Utilizing this resource, we extract diagnoses and develop a pipeline for symptom extraction from the free text in Spanish. As EHR are not primarily designed for research, their use depends on a series of steps that we have undertaken to ensure that their information is valid for this purpose[18]. First, we estimated the reliability of diagnoses. Second, we evaluated automated symptom extraction in individual sentences as well as

lifetime, compared to chart review. We further verified the alignment between symptoms and diagnostic specifiers at individual visits (e.g., we expect symptoms to be less frequent during visits labeled as *mild* or *in remission* compared to those labeled as *severe*).

Having established the high quality of extracted phenotypes from EHR, we then investigated disease trajectories. We characterized both diagnostic switches and the accumulation of comorbidities in SMI patients. We quantified, and identified modifiers of, the diagnostic stability of SMI; in particular, we evaluated whether specific symptoms predict future diagnostic changes.

Methods

EHR database

Until 2019, the CSJDM was the only mental health care facility in the Caldas region of Colombia[17]. Its electronic health records (EHRs), utilized for recording all patient visits since 2005, provide a comprehensive database of diagnosed mental illness for patients from the whole region during this period. The EHRs contain both structured and unstructured data. Structured data include demographic information (e.g., age and gender), diagnostic codes (ICD-10), vital signs, medication, and healthcare system use (e.g., duration and type of visits: inpatient, outpatient, emergency room). Unstructured data include intake, discharge, and progress notes from both inpatient and outpatient visits as well as emergency room notes. Using procedures approved by the IRBs from UCLA and CSJDM, prior to the analyses reported here, we removed from the EHRs any fields considered Protected Health Information by HIPAA [19]. In addition, names and numbers exceeding 5-digits (potential ID numbers) were stripped from the text using regular expression matching.

For our analyses, we included all patients in CSJDM with at least one clinical note (n=77,538) and excluded patients with missing gender information (n=626). We further excluded visits outside the age range of 4-90 or with errors in diagnostic coding (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 physician. Over time this procedure generates a time-stamped diagnostic sequence. We extracted this sequence for every patient and selected those who had at least one primary diagnosis of Severe Mental Illness (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 has 22,447 patients with 157,003 visits (Supplementary Figure 1).

Primary diagnosis classification and reliability estimation

As a first step in determining the reliability of the ICD-10 diagnoses, we assessed the agreement of diagnostic codes with those obtained through a complete chart review. A research clinician (MC) with 10+ years of experience, reviewed charts from 120 patients (40 each with ICD-10 codes of MDD (F32X/F33X), BD (F31X), and SCZ (F20X)), whom we selected at random from a large ongoing CSJDM study of individuals with SMI [20]. The clinician filled out an item checklist and determined the *current* primary diagnosis based on recorded DSM-5 criteria [21]. We evaluated the agreement between this diagnosis and the *most recently* recorded SMI code using Cohen's kappa [22]. A total of 15 participants did not have an SMI code at the time of review and were excluded from this analysis resulting in n=105.

NLP algorithm to extract clinical features

To extract clinical features, such as symptoms and behaviors, from clinical notes written in Spanish, we developed a Natural Language Processing (NLP) algorithm with two components: Named Entity Recognition and Negation Detection (for details see Supplementary Note 1). To train the algorithm, two clinicians annotated a corpus of 3,600 sentences, identifying syntactic patterns representing target features. For this report, we focused specifically on four clinical features: Suicide Attempts, Suicidal Ideation, Delusions, and Hallucinations (Supplementary Table 2) and evaluated them using a held-out set of 290 gold standard sentences and reported precision, recall, and their harmonic mean (F1). To validate the NLP features at the individual patient level, we contrasted items from the chart review checklist with an NLP-based "lifetime" phenotype, defined as having two or more notes with affirmative mentions of a specific feature.

Visit-level data validation: NLP features and ICD-10 diagnoses

We validated features cross-sectionally by their association with fine-grained ICD-10 codes *assigned at individual visits* in the SMI cohort (n=22,447). Each visit is labeled with a three-digit ICD-10 code representing the *current state* of the patient. In BD, for example, F31X can represent an episode of hypomania (X=0), mania with psychotic features (X=2), mild depression (X=3), remission (X=7), etc. (Supplementary Table 1). Valid features should corroborate these specifiers. For example, we expect psychotic symptoms (Delusions and Hallucinations) to be noted commonly during visits labeled as "*with psychotic features*", but only rarely during visits labeled as *mild* or *in remission*. We additionally tested the hypothesis that, in BD, Suicidal Ideation and Suicide Attempts are present to a greater extent in

depressed episodes compared to manic episodes [23]. For MDD, BD, and SCZ, we tested the association between ICD-10 code specifiers and the presence of NLP features. Using a mixed-effect logistic regression we correct for confounding factors such as inpatient status and account for multiple visits per person. Aside from validating feature endorsement *at specific visits*, we tested whether NLP features can *anticipate* diagnostic changes in future visits. Specifically, we tested whether the psychosis features predict the application of psychosis specifiers in ICD-10 mood disorders designations. (Supplementary Note 2)

Patient-level data validation: NLP features and ICD-10 diagnoses

To evaluate the reliability of *lifetime* phenotypes, we tested for associations between the presence of NLP features and diagnoses of MDD, BD, and SCZ, at the *individual level*. Association tests were performed using logistic regression, adjusting for length of record and stratifying by hospitalization history (Supplementary Note 3).

Diagnostic switches, comorbidities, and trajectories

We defined two types of diagnostic changes: diagnostic switches and comorbidities. Here, we operationalize diagnostic switches as changes between two psychiatric diagnoses that cannot, by definition, be held at the same time; here we use this term to refer to the diagnoses in the ICD-10 F3 and F2 chapters (F3XX and F2XX respectively - excluding F32; see Supplementary Table 3 and Supplementary Note 4 for details). Comorbidities, by contrast, refer to all other combinations of ICD-10 codes; comorbid diagnoses accumulate over time, without limit. Using these definitions, an individual patient's diagnostic trajectory may include 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. For individual SMI diagnoses, we estimated long-term prospective and retrospective stability from patient trajectories with 10 or more visits [24]. 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 30) were evaluated using z-tests (Supplementary Note 5).

We explored factors contributing to visit-to-visit diagnostic stability, as follows. First, we used a mixed-effect logistic regression to estimate how time (measured by visit number) affects the probability

of switching diagnoses, adjusting for repeated patient observations. Then, we expanded this model to evaluate the effects of patient sex and age, as well as those of primary diagnosis, inpatient status, previous switch, NLP features, and receiving a Not Otherwise Specified (NOS) code at the previous visit (Supplementary Note 6). An NOS diagnosis is used to express uncertainty in cases of atypical or confusing patient presentation, or when temporal criteria are not yet met [25], and therefore serves as a positive control. In sensitivity analysis, we measured time in years since the first encounter rather than by visit number.

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

EHR-extracted diagnoses and features show high agreement with the chart review

As of June 2022, the CSJDM EHR database included 303,762 visits of 71,856 patients aged 4-90 (50.8% female); 58,502 visits of 30,729 patients were for in-patient hospitalizations, while the remaining 245,260 visits (58,262 patients) were for outpatient care. 22,447 patients had received an SMI diagnosis at some point in their trajectory (Table 1). The diagnoses obtained for the 105 patients for whom we conducted a complete manual chart review, demonstrated "very good" to "excellent" agreement [26] with those obtained from the most recent ICD-10 code alone. Kappa estimates for specific diagnoses 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). Estimates were similarly high when considering inpatient visits only (Supplementary Table 4).

Validation of NLP features. Annotation was consistent between clinicians, with inter-annotator agreement ranging from "good" to "excellent" (Supplementary Table 2). Feature extraction at the sentence level showed excellent performance (precision range: 0.88-1.0 and recall range: 0.62-1.0; Supplementary Table 5), resulting in a satisfactory F1 for all features (Suicide Attempt: 0.82, Suicidal Ideation: 0.73, Delusions: 1.0, Hallucinations: 0.95, see Supplementary Note 7 for a description of errors). At the patient level, two separate affirmative mentions of NLP-extracted features demonstrated high concordance with lifetime phenotypes from manual chart review (F1 values for Suicide Attempt: 0.72, Suicidal Ideation: 0.80, Delusions: 0.82 and Hallucinations: 0.81; Supplementary Table 5). As expected, requiring more mentions for a positive classification increased precision, at the cost of decreased recall;

an optimal balance (as measured by the F1 statistic) was achieved by requiring affirmative mentions in at least two notes (Supplementary Figure 2).

In a post-hoc review of true and false positive patient records (Supplementary Note 1), both clinicians identified the feature as being present in all but one of the true positives and marked about half of the previously identified false positives as true, effectively increasing the observed F1: Suicide Attempt: 0.77, Suicidal Ideation: 0.90, Delusions: 0.85 and Hallucinations: 0.86 (Supplementary Table 5).

NLP Feature profiles align with specific states of major mood disorders and schizophrenia

At the visit level, feature profiles extracted from free-text notes closely align with ICD-10 codes (Figure 1; Supplementary Table 6). For example, we verify that, in MDD visits, Delusions and Hallucinations are more common in severe episodes denoted as "with psychotic features" than in similar episodes without such a designation (ORs: 14.8 and 13.9, respectively, p-values $<1e^{-50}$). Similarly, very few affirmed instances (<3%) were extracted for visits recorded as "Mild" depressive episodes, or for BD or MDD "In remission". As hypothesized, among patients with diagnoses of BP, visits recorded as being for manic episodes had lower odds of mentioning Suicide Attempt and Suicidal Ideation compared to depressive episodes (ORs 0.12 and 0.07, respectively; p-values $<1e^{-7}$).

We extracted an unexpectedly high frequency of Delusions in visits recorded as being for mania, even if these visits were recorded as "mania without psychotic features" (F311; 46.6%). The inclusion of Delusions classically considered part of mania (those involving grandiosity) only partially explain this observation (up to 32% of instances, Supplementary Figure 3, see Supplementary Note 8 for additional observations).

We hypothesized that features observed at a given visit could *anticipate* ICD-10 code changes in the next visit. Indeed, during severe episodes of mood disorders, the presence of psychotic features mentioned in the text significantly increased the odds of a psychotic diagnosis in the following visit (OR: 1.52; p-value= $3e^{-5}$). The same pattern holds for the special case of BD visits with the ICD code "Manic episode without psychotic symptoms" (F311); when followed by an ICD code that includes psychotic features (e.g., F312), psychosis is much more likely identified in the text than when followed by an ICD code that does not include psychotic features (68% vs 47%, p-value $1e^{-7}$).

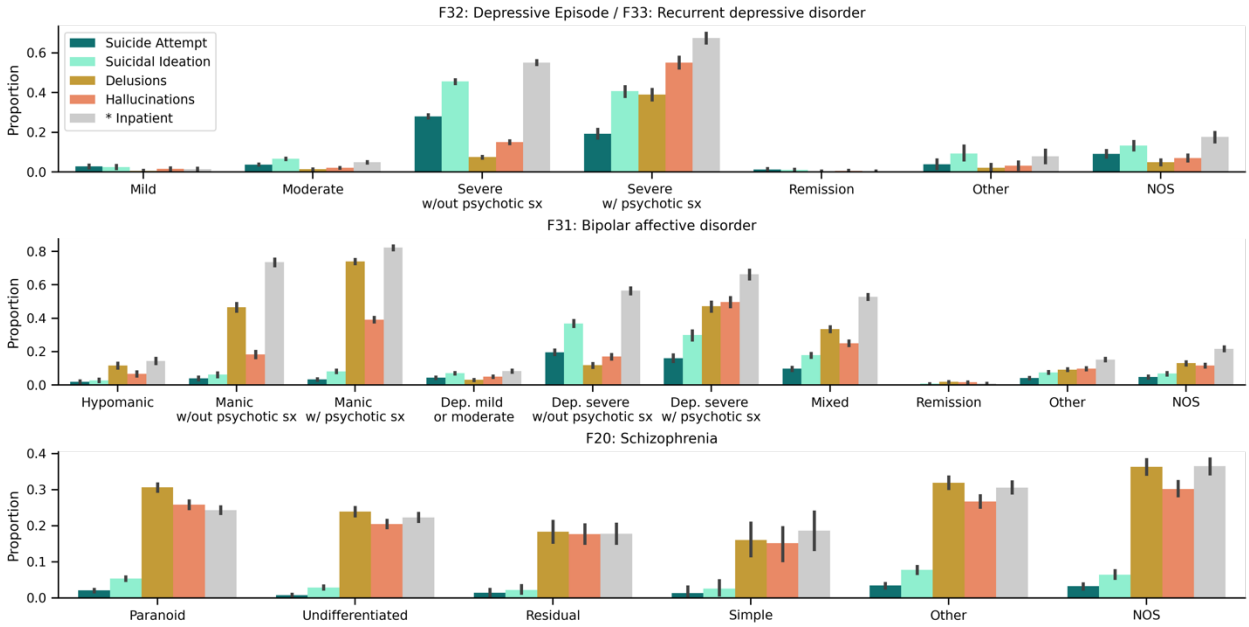

Figure 1: Association between ICD codes and NLP features recorded within a visit. The proportion of inpatient visits for each diagnostic code is shown in gray. Error bars indicate 95% confidence intervals obtained through bootstrapping. Individual patients may be present more than once in each specific state. N=22,447 patients.

Clinical and demographic characteristics of SMI patients

Using the framework we developed for extracting clinical phenotypes, we observe that demographic characteristics, indices of overall clinical course, and NLP features across MDD, BD, and SCZ broadly align with expectations from the literature (Table 1 and Supplementary Figure 4): The majority of patients with MDD and BD but not SCZ are female (66.4% and 65.1% vs. 24.3% respectively) [27,28]; patients with SCZ and BD diagnoses, compared to MDD, are hospitalized significantly more frequently (74.7% and 71.4% vs 56.4% respectively) and have longer hospital stays (median of 14 and 12 vs 8 days) [29] (see Supplementary Table 7 for details on statistical tests performed).

The lifetime prevalence of NLP features varies between diagnoses, even after accounting for the number of notes recorded for a patient. For example, in previously hospitalized patients, Suicide Attempt and Suicidal Ideation are less frequent in BD than in MDD (ORs 0.41 and 0.32 respectively; p-values $<1e^{-80}$) but more frequent than in SCZ (ORs 2.05 and 1.91 respectively; p-values $<1e^{-20}$). Irrespective of patients' hospitalization history, Delusions and Hallucinations are most commonly recorded in SCZ patients (78.4% and 75.8%; [21]), followed by BD (52.3% and 48.4%), and MDD (16.2% and 25.1%, BD-MDD OR range 1.47-4.64; p-values $<1e^{-4}$; BD-SCZ OR range 0.11-0.30; p-values $<3e^{-28}$; Supplementary Table 7).

	Major Depressive Disorder	Bipolar Disorder	Schizophrenia	All
Patients, N	11051	8325	2771	22447
Visits, N	47783	78877	27593	157003
Visits per patient, N	2 (1, 5)	5 (2, 12)	6 (2, 13)	3 (1, 9)
Any inpatient visits, %	56.4	71.4	74.7	64.4
Any ER visits, %	54.0	56.5	62.4	55.8
Age at most recent visit, years	37.7 (21.6, 54.8)	45.1 (27.4, 58.5)	33.7 (23.6, 52.0)	40.0 (24.1, 56.2)
Female, %	66.4	65.1	24.3	60.4
Any visits under age 18, %	19.7	12.2	17.1	16.4
Length of stay, days	8 (4, 12)	12 (7, 18)	14 (9, 21)	11 (6, 17)
Length of medical record, years	7.0 (4.1, 10.7)	10.2 (6.4, 14.1)	9.0 (5.8, 14.2)	8.4 (5.1, 12.8)
Suicide Attempt (out, in), %	33.7 (10.3, 47.3)	23.0 (9.3, 27.6)	13.4 (2.2, 16.4)	26.8 (9.2, 34.3)
Suicidal Ideation (out, in), %	53.0 (19.8, 72.2)	40.3 (13.9, 49.0)	29.0 (6.7, 35.0)	44.8 (16.7, 56.9)
Delusions (out, in), %	16.2 (3.8, 23.4)	52.3 (10.6, 66.1)	78.4 (42.8, 87.9)	38.9 (10.4, 51.2)
Hallucinations (out, in), %	25.1 (6.8, 35.8)	48.4 (13.4, 60.0)	75.8 (42.5, 84.6)	41.0 (12.6, 53.3)

Table 1. Demographic and clinical characteristics of the study cohort. Patients are classified by their most recent SMI diagnosis. Medians with interquartile range (IQR) are presented for: visits per patient, age at the most recent visit, length of stay, and length of the medical record. NLP features show the percentage of patients with a 'lifetime' endorsement (affirmative mentions in at least two notes). For all cells in which multiple percentages are shown, the initial value refers to those observed in all patients, while the values in parentheses refer to those with only outpatient visits and those with at least one inpatient visit, respectively. Statistics comparing different diagnoses are provided in Supplementary Table 7

The diversity of diagnostic trajectories underlies the low retrospective stability of SMI

The majority of SMI patients with at least three visits (n=12,962) have multiple diagnoses recorded in their EHR (64% of patients, Figure 2A). Specifically, 20% of SMI patients display diagnostic switches (teal bars), 30% display comorbidities (orange bars; Supplementary Table 8), and 15% display both switches and comorbidities (purple bars).

While some trajectories are common, for example, the switch from MDD to BD (observed in 22% of current BD patients) 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.

The level of diagnostic instability observed for different diagnoses accords with what would be expected from their characteristic trajectories (Figure 2B). We estimated prospective and retrospective stability, for each diagnosis, evaluating trajectories of 10 or more visits (n=5,016). Prospective stability is lower for MDD than BD or SCZ (56% vs. 88% and 83%, respectively; 2-df chi-square= 383; p-value $5e^{-84}$). Retrospective stability, by contrast, while low for all diagnoses, is highest in MDD (53% vs. 48% and 40% in BD and SCZ respectively; 2-df chi-square= 34.5; p-value $3e^{-8}$). (Supplementary Note 9 and Supplementary Table 9).

The four NLP-derived features provide layers of trajectory information beyond those available from the evaluation of diagnoses. They subdivide the patient population into 16 groups (Figure 2C). As expected, the presence of a feature during one visit increases the probability of observing the same feature on the next visit, but patterns of cross-feature probabilities are harder to explain. For example, the probability of extracting Delusions after having extracted Hallucinations is higher than the converse (P(Hal->Del):0.31; P(Del->Hal):0.26; p-value $1e^{-21}$). (Figure 2D).

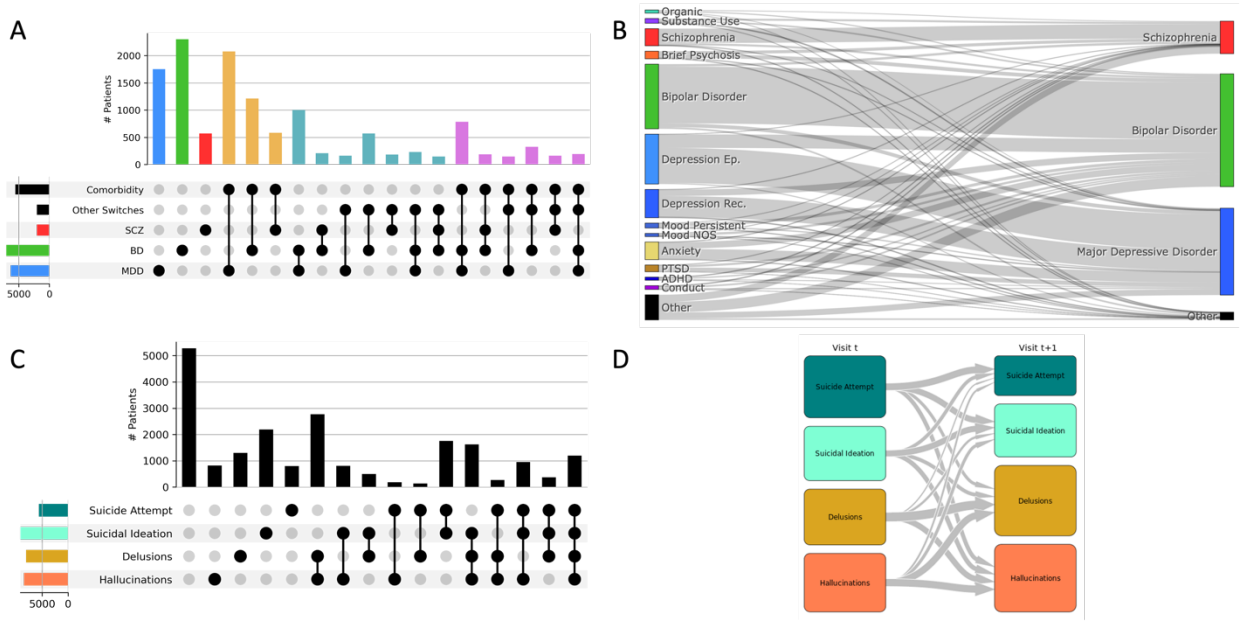

Figure 2: Disease trajectories of SMI in patients with at least three visits. A) UpSet plot presenting diagnostic switches and comorbidities. Patients with a single SMI diagnosis (blue, green, red, total n=4,633); a single SMI diagnosis and other comorbidities (orange n=3,878); multiple SMI diagnoses and no other comorbidities (teal n=2,530); multiple SMI diagnoses and other comorbidities (purple, n=1,921). Bars with n<75 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 in SMI are shown in Supplementary Figure 5). C) Combinations of NLP features extracted from each patient in their complete medical record. D) Feature transitions. Arrows represent the probability of recording an NLP feature in the next visit, given the presence of a feature in the current visit.

Time, previous history, and other factors affecting diagnostic stability

Diagnostic switching is most frequent during the early stages of treatment. While 11.3% of the patients change diagnosis on their second visit, this percentage decreases over time (Figure 3A; log10(k) OR 0.56, p-value $5e^{-66}$) and stabilizes at around 4% after the tenth visit. Additional predictors of diagnostic instability include: a diagnostic switch in the previous visit (Figure 3B; 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, of the NLP features, Delusions and Hallucinations (OR 1.50 and 1.17, p-values $2e^{-18}$ and $3e^{-4}$, respectively). On the other hand, diagnostic stability is associated with diagnoses of SCZ or BD compared to MDD (OR 0.31 and 0.32; p-values $<3e^{-70}$), male gender (OR 0.71, p-value $2e^{-16}$), and age (OR per decade= 0.96, p-value $8e^{-4}$). Sensitivity analyses confirmed these findings; the same pattern is observed if we model switching by time rather than visit number (Supplementary Figure 6).

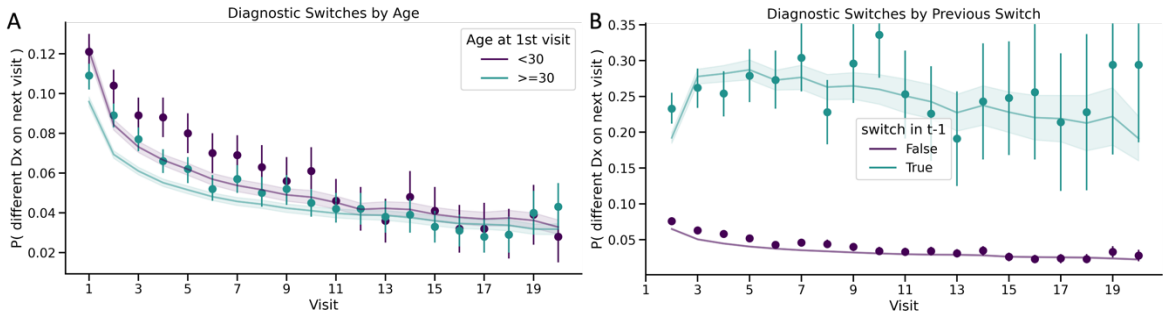

Figure 3: Diagnostic stability over time. At each visit k, the proportion of patients that will switch primary diagnosis code on their next visit k+1. A) Stratified by age groups: before and after 30. B) Stratified by having switched diagnoses from visit k-1. N=12,962 patients.

Discussion

Analyses of longitudinal EHR data collected over 17 years for more than 20,000 patients with SMI reveal a large diversity of disease trajectories with frequent diagnostic switching. Incorporating diagnostic codes with NLP-extracted features from clinical notes, we investigated both fine-grained visit-to-visit disease progression and long-term illness trajectories, quantifying diagnostic instability and factors affecting it. Our validation of diagnostic codes in the EHR suggests that diagnostic switching reflects true diagnostic instability in psychiatric disease.

High quality of diagnoses and clinical features

We evaluated the quality of diagnostic codes recorded in the EHR through their concordance with chart review and with other clinical features extracted from the EHR, including from clinical notes written in Spanish. The reliability of current diagnoses is high, with kappa values superior to those from similar studies[18], and comparable to those reported by Castro et al. [30] based on a large hospital database.

Automated extraction of clinical features from Spanish text is still in its infancy [31,32]. Due to high variability in vocabulary and language across healthcare settings, 'off-the-shelf' NLP pipelines often have poor performance [33,34], present serious usability limitations [35], or have limited sentence-level recall [36]. We, therefore, designed a custom rule-based pipeline and showed that it reliably extracts clinical features, such as symptoms and behaviors, from narrative notes in our data. This pipeline is the first of its kind used for psychiatric notes in Spanish.

We showed that two distinct affirmative mentions of a feature optimize the balance between precision and recall. Depending on the analysis goals, however, one may modify this threshold to favor precision, (increasing the confidence that a patient truly presents with a symptom), or recall ("casting a wider net"). Importantly, our algorithm identified clinical features in the records of multiple individuals where the initial chart review failed to do so, highlighting the limitations of manually reviewing large numbers of notes. This observation suggests that tools that provide clinicians with an automated way of reviewing specific features in a patient's notes could be useful in clinical care.

Diverse diagnostic trajectories

Using data from a psychiatric secondary care facility, our study is the first to combine structured and unstructured EHR data to comprehensively characterize longitudinal diagnostic trajectories of SMI. Most prior studies, conducted in large healthcare systems, aimed either at identifying psychiatric cohorts from a wider patient population [30] or at predicting specific outcomes, such as first-episode psychosis [37] and suicidal behavior [38].

A novelty of our approach is that, using administrative data, we draw distinctions between the accumulation of comorbidities and diagnostic switching. We show that about half of the SMI population has accumulated one or more psychiatric comorbidities and that over one-third has switched diagnoses. Given our chart-review validation of current diagnoses, these switches are unlikely to be artifacts of code assignment. These switches generate a broad variety of disease trajectories, showing that most patients exhibit different, rare trajectories.

Patient presentation is expected to be most uncertain and dynamic during the early stages of illness [14,39]. While we observe that diagnostic switching is most common during patients' first few visits

to the clinic, it is not limited to these early encounters and frequently persists well into their trajectories. Even though instability decreases with age, time in treatment may be a more important driver of diagnostic stability of SMI.

Our finding that a current switch increases the probability of a future switch is driven in large part by a small number of patients who rapidly accumulate diagnoses. These individuals may constitute a group for whom our current diagnostic system is particularly poorly suited to properly describe their illness, but who may not be identified in research based on cross-sectional assessments. By contrast, in EHR-based studies it is straightforward to flag such individuals, creating opportunities to discover distinct features that characterize this subgroup and possibly reduce the heterogeneity of the different diagnostic groups in the overall dataset.

The EHR database also enables the identification of trajectories based on NLP-derived features from clinical notes, rather than on diagnoses alone. Notably, recorded changes in psychosis features anticipate future psychosis diagnoses. This observation supports the view that symptom-level analyses of EHR text will prove a particularly informative approach for delineating SMI trajectories.

Limitations

From an algorithmic standpoint, the feature extraction method presented here is a proof-of-concept; while currently limited to a handful of symptoms and behaviors, it can be expanded to additional phenotypes. Furthermore, contextual information captured by our method is currently limited to negations. Future improvements will focus on capturing richer contexts, such as person (e.g., “my mother had delusions”) or hypothetical futures (e.g., “suspend medication in case of hallucinations”).

Recommendations and future directions

Our results support the view that research classifications that incorporate past and future trajectory data will likely be much less heterogeneous and more realistic than current systems that classify patients according to a single ‘lifetime’ diagnosis. Because datasets with thousands of uncommon trajectories will be infeasible to analyze, an important focus of future work should be the development of better approaches to reduce dimensionality by clustering patients with similar trajectories[40]. EHR databases usually include information on interventions, such as pharmacological treatments, that may influence disease trajectories. An important direction of future research will be to model this impact.

Conclusions

We showed that high-quality and highly granular EHRs can be used to characterize disease trajectories of SMI. EHR databases enable longitudinal research studies in LMICs that have previously been feasible only in UICs, using data from patient registries and longitudinal cohorts

Acknowledgments

Research reported here was supported by R01MH123157 (to LMOL, CLJ, and NBF), R01MH113078 (to CEB, CLJ, and NBF), R00MH116115 (to LMOL), T32MH073526 (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 necessarily represent the official views of Fulbright or the National Institutes of Health.

References

1. Siggaard, T. et al. Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million Danish patients. Nat. Commun. 11, 4952 (2020).
2. Regier, D. A. et al. DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. Am. J. Psychiatry 170, 59–70 (2013).
3. Hyman, S. E. The diagnosis of mental disorders: the problem of reification. Annu. Rev. Clin. Psychol. 6, 155–179 (2010).
4. Kendler, K. S. The nature of psychiatric disorders. World Psychiatry 15, 5–12 (2016).
5. Psychiatrie, Ein Kurzes Lehrbuch für Studirende und Aerzte. Von Dr. Emil Kraepelin, Professor in Dorpat. Zweite, gründlich umgearbeitete Auflage. Leipzig: Verlag von Ambr. Abel. 1887. Journal of Mental Science vol. 33 601–601 (1888).
6. Fritz, K. et al. Is a delay in the diagnosis of bipolar disorder inevitable? Bipolar Disorders vol. 19 396–400 (2017).
7. Grande, I., Berk, M., Birmaher, B. & Vieta, E. Bipolar disorder. The Lancet vol. 387 1561–1572 (2016).
8. Musliner, K. L. & Østergaard, S. D. Patterns and predictors of conversion to bipolar disorder in 91 587 individuals diagnosed with unipolar depression. Acta Psychiatr. Scand. 137, 422–432 (2018).
9. Kessler, R. C. et al. Anxious and non-anxious major depressive disorder in the World Health Organization World Mental Health Surveys. Epidemiol. Psychiatr. Sci. 24, 210–226 (2015).
10. Kalman, J. L. et al. Characterisation of age and polarity at onset in bipolar disorder. Br. J. Psychiatry 219, 659–669 (2021).
11. 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 vol. 177 155–163 (2020).

12. Musliner, K. L. et al. Polygenic Risk and Progression to Bipolar or Psychotic Disorders Among Individuals Diagnosed With Unipolar Depression in Early Life. Am. J. Psychiatry 177, 936–943 (2020).
13. Jonas, K. G. et al. Schizophrenia polygenic risk score and 20-year course of illness in psychotic disorders. Transl. Psychiatry 9, 300 (2019).
14. Plana-Ripoll, O. et al. Exploring Comorbidity Within Mental Disorders Among a Danish National Population. JAMA Psychiatry 76, 259–270 (2019).
15. 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 (2022) doi:10.1001/jamapsychiatry.2022.0685.
16. 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).
17. Song, J. et al. Geospatial analysis reveals distinct hotspots of severe mental illness. doi:10.1101/2022.03.23.22272776.
18. 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, 263 (2016).
19. Office for Civil Rights (OCR). Methods for De-identification of PHI. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html (2012).
20. 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. Lancet Psychiatry 7, 411–419 (2020).
21. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). (American Psychiatric Pub, 2013).
22. Cohen, J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement vol. 20 37–46 (1960).
23. Dome, P., Rihmer, Z. & Gonda, X. Suicide Risk in Bipolar Disorder: A Brief Review. Medicina 55, (2019).
24. Schwartz, J. E. et al. Congruence of diagnoses 2 years after a first-admission diagnosis of psychosis. Arch. Gen. Psychiatry 57, 593–600 (2000).
25. 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).
26. 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. Am. J. Psychiatry 170, 43–58 (2013).
27. Salk, R. H., Hyde, J. S. & Abramson, L. Y. Gender differences in depression in representative national samples: Meta-analyses of diagnoses and symptoms. Psychol. Bull. 143, 783–822 (2017).

28. Jongsma, H. E., Turner, C., Kirkbride, J. B. & Jones, P. B. International incidence of psychotic disorders, 2002-17: a systematic review and meta-analysis. Lancet Public Health 4, e229–e244 (2019).
29. Crossley, N. E. & Sweeney, B. Patient and Service-Level Factors Affecting Length of Inpatient Stay in an Acute Mental Health Service: A Retrospective Case Cohort Study. doi:10.21203/rs.3.rs-40278/v2.
30. Castro, V. M. et al. Validation of electronic health record phenotyping of bipolar disorder cases and controls. Am. J. Psychiatry 172, 363–372 (2015).
31. Akhtyamova, L. Named Entity Recognition in Spanish Biomedical Literature: Short Review and Bert Model. 2020 26th Conference of Open Innovations Association (FRUCT) (2020) doi:10.23919/fruct48808.2020.9087359.
32. Cotik, V., Rodríguez, H. & Vivaldi, J. Spanish Named Entity Recognition in the Biomedical Domain. Information Management and Big Data 233–248 (2019) doi:10.1007/978-3-030-11680-4_23.
33. AlShuweihi, M., Salloum, S. A. & Shaalan, K. Biomedical Corpora and Natural Language Processing on Clinical Text in Languages Other Than English: A Systematic Review. Studies in Systems, Decision and Control 491–509 (2021) doi:10.1007/978-3-030-47411-9_27.
34. Névéol, A., Dalianis, H., Velupillai, S., Savova, G. & Zweigenbaum, P. Clinical Natural Language Processing in languages other than English: opportunities and challenges. J. Biomed. Semantics 9, 12 (2018).
35. Zheng, K. et al. Ease of adoption of clinical natural language processing software: An evaluation of five systems. J. Biomed. Inform. 58 Suppl, S189–S196 (2015).
36. Jackson, R. G. et al. Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. BMJ Open 7, e012012 (2017).
37. 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).
38. Barak-Corren, Y. et al. Predicting Suicidal Behavior From Longitudinal Electronic Health Records. Am. J. Psychiatry 174, 154–162 (2017).
39. Bromet, E. J. et al. Diagnostic shifts during the decade following first admission for psychosis. Am. J. Psychiatry 168, 1186–1194 (2011).
40. Krebs, M. D. et al. Associations between patterns in comorbid diagnostic trajectories of individuals with schizophrenia and etiological factors. Nat. Commun. 12, 6617 (2021).