Journal of Affective Disorders 356 (2024) 64–70

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Research paper

Identifying data-driven subtypes of major depressive disorder with electronic health records

Abhishek Sharma [b], Pilar F. Verhaak [a], Thomas H. McCoy [a,c], Roy H. Perlis [a,c,*], Finale Doshi-Velez [b,**]

[a] Center for Quantitative Health, Massachusetts General Hospital, 185 Cambridge Street, Boston, MA, 02114, United States of America
[b] Harvard John A. Paulson School of Engineering and Applied Sciences, 29 Oxford Street, Cambridge, MA 02138, United States of America
[c] Harvard Medical School, 25 Shattuck Street, Boston, MA 02115, United States of America

ARTICLE INFO

Keywords:
Heterogeneity
Major depressive disorder
Representation learning
Latent Dirichlet allocation
Data-driven subtypes

ABSTRACT

Background: Efforts to reduce the heterogeneity of major depressive disorder (MDD) by identifying subtypes have not yet facilitated treatment personalization or investigation of biology, so novel approaches merit consideration.
Methods: We utilized electronic health records drawn from 2 academic medical centers and affiliated health systems in Massachusetts to identify data-driven subtypes of MDD, characterizing sociodemographic features, comorbid diagnoses, and treatment patterns. We applied Latent Dirichlet Allocation (LDA) to summarize diagnostic codes followed by agglomerative clustering to define patient subgroups.
Results: Among 136,371 patients (95,034 women [70 %]; 41,337 men [30 %]; mean [SD] age, 47.0 [14.0] years), the 15 putative MDD subtypes were characterized by comorbidities and distinct patterns in medication use. There was substantial variation in rates of selective serotonin reuptake inhibitor (SSRI) use (from a low of 62 % to a high of 78 %) and selective norepinephrine reuptake inhibitor (SNRI) use (from 4 % to 21 %).
Limitations: Electronic health records lack reliable symptom-level data, so we cannot examine the extent to which subtypes might differ in clinical presentation or symptom dimensions.
Conclusion: These data-driven subtypes, drawing on representative clinical cohorts, merit further investigation for their utility in identifying more homogeneous patient populations for basic as well as clinical investigation.

1. Introduction

For many decades, recognizing the heterogeneity of clinical presentations and treatment responses in major depressive disorder, there have been efforts to identify depressive subtypes that might yield more homogeneous groups. Early efforts relied on responsiveness to monoamine oxidase inhibitors (MAOIs) to define atypical and melancholic depression (Hyman Rapaport, 2007); these distinctions became less useful with the failure to identify associated biomarkers and the recognition that these groups did not differ in response to newer antidepressants (ADs). Subsequent work focused on symptom profiles or psychiatric comorbidities (Fava et al., 2006), including anxiety (Stavrakaki and Vargo, 1986), and irritability (Perlis et al., 2005).

All of these efforts relied on characteristics identified a priori by investigators, albeit on the basis of clinical experience. An alternative approach to defining subgroups relies on data-driven or unsupervised strategies, allowing discovery of novel subgroups based on patterns in patient data. One such study attempted to identify symptom-based subtypes of depression (van Loo et al., 2012), for example. However, this data-driven strategy has proven to be challenging because most treatment studies, the source until recently of most phenotypic data, exclude patients with medical and psychiatric comorbidities (Rush et al., 2004). As such, clinical trials did not reflect the range of variation in the general population that might be necessary to adequately capture subgroups.

A range of strategies has been employed to identify novel phenotypes from electronic health records. One recent investigation applied autoencoders as a means of feature engineering (Jones et al., 2023); while this strategy captures temporal relationships that would be missed by clustering, these features did not meaningfully improve predictions. Another strategy examined polygenic liability for depression in electronic health records, in an effort to capture systemic features that may

* Correspondence to: R.H. Perlis, Massachusetts General Hospital, 185 Cambridge Street, 6th Floor, Boston, MA 02114, United States of America.
** Correspondence to: F. Doshi-Velez, Harvard University, 150 Western Avenue, Allston, MA 02134, United States of America.
E-mail addresses: rperlis@mgh.harvard.edu (R.H. Perlis), finale@seas.harvard.edu (F. Doshi-Velez).

https://doi.org/10.1016/j.jad.2024.03.162
Received 4 December 2023; Received in revised form 25 March 2024; Accepted 26 March 2024
Available online 1 April 2024

share biology with major depression (Fang et al., 2022). Other recent work has applied natural language processing (NLP) to individuals with depressive disorders in electronic health records, for example to capture features of mania (Patel et al., 2022) or dimensional measures of agitation (Hart et al., 2021). However, these latter approaches require additional data types that may not be available in large representative health systems, and sometimes rely on narrative notes that transfer poorly across health systems.

As an alternative, to better understand the variability of depressive phenotypes in less selected clinical populations, we drew on coded clinical data in electronic health records of 2 large academic medical centers and their affiliated community hospitals and outpatient clinics. We hypothesized that coded clinical data would identify recognizable clinical subpopulations, and potentially define new subtypes for further investigation.

2. Methods

2.1. Data sources and input features

We drew on electronic health records from two large academic medical centers and their affiliated community hospitals and outpatient clinics. Available patient data included diagnostic codes, including ICD-9 and ICD-10 codes, CPT codes capturing lab tests and procedures, and RXNORM codes capturing medications along with the timestamps of the codes. These data also included sociodemographic features including age, self-reported gender, race, and ethnicity.

The study cohort included patients with at least one diagnosis of major depressive disorder between 2017 and 2022 based on ICD-9 or ICD-10 codes (see Fig. 1 for patient demographic information and S1 for all codes). Patients without an AD RXNORM code were excluded in addition to those patients who were younger than 18 years or older than 80 years at their last code. Only patients with codes observed both before and after AD prescription were included. For a full description of cohort derivation, please reference Fig. 2.

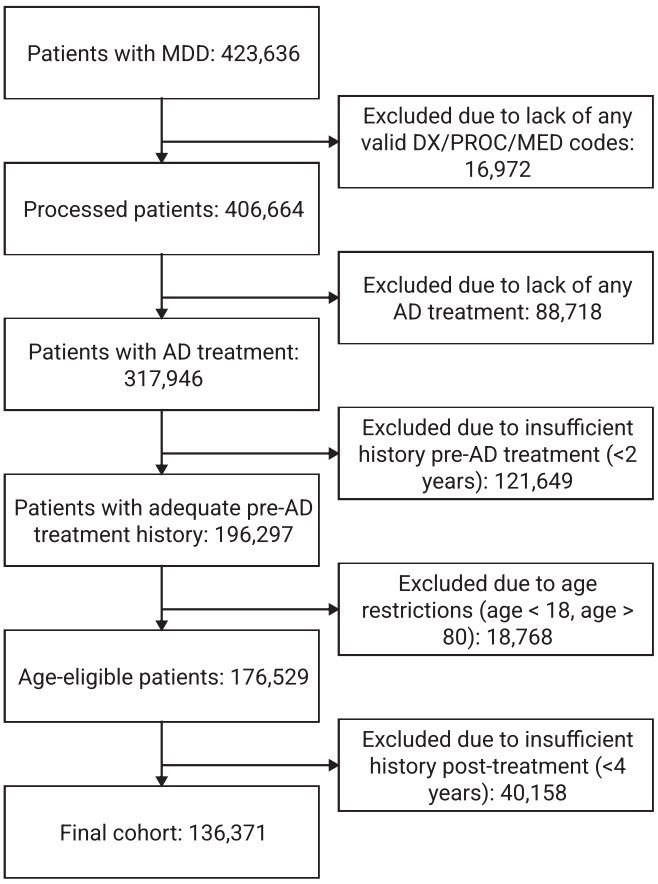

Fig. 2. CONSORT flow diagram of cohort derivation.

Patient Demographics

	Cluster 0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
Count	4,744	13,466	7,959	4,155	3,232	12,619	3,810	5,499	4,981	38,701	2,577	12,993	6,889	5,816	8,930	136,371
Age: 18-30	38%	32%	8%	53%	32%	10%	17%	17%	2%	9%	6%	20%	5%	1%	2%	15%
Age: 30-40	53%	35%	11%	33%	24%	13%	25%	26%	8%	16%	7%	22%	12%	6%	7%	19%
Age: 40-50	9%	23%	21%	10%	18%	20%	28%	26%	21%	23%	16%	22%	26%	33%	22%	22%
Age: 50-60	0%	8%	28%	3%	15%	26%	20%	19%	32%	25%	25%	19%	30%	39%	33%	23%
Age: 60-70	0%	2%	23%	1%	8%	21%	8%	9%	27%	19%	27%	12%	19%	18%	25%	15%
Age: 70-80	0%	0%	8%	0%	3%	9%	2%	3%	10%	7%	19%	4%	8%	4%	10%	6%
Gender: F	100%	96%	68%	92%	58%	66%	70%	61%	78%	64%	58%	51%	66%	99%	53%	70%
Gender: M	0%	4%	32%	8%	42%	34%	30%	39%	22%	36%	42%	49%	34%	1%	47%	30%
Race: Asian	3%	2%	2%	3%	4%	2%	2%	2%	2%	2%	2%	1%	1%	2%	2%	2%
Race: Black	11%	10%	3%	13%	6%	7%	5%	4%	3%	6%	4%	9%	7%	3%	6%	7%
Race: White	59%	63%	90%	62%	78%	79%	78%	86%	90%	82%	87%	79%	79%	86%	78%	79%
Race: Other	27%	25%	6%	21%	13%	12%	15%	8%	5%	10%	8%	11%	13%	9%	14%	13%
Ethnicity: Hispanic	11%	13%	2%	6%	5%	6%	7%	3%	2%	4%	3%	4%	6%	5%	8%	6%

Fig. 1. Patient demographic information.

The Massachusetts General Brigham HealthCare institutional review board approved the study protocol, waiving the requirement for informed consent since only deidentified data were used and no participant contact was required.

2.2. Exclusion criteria

The initial cohort was defined as patients with at least one MDD diagnosis defined by ICD code ($N = 423{,}636$). We excluded patients without any AD RXNORM code ($N = 88{,}718$). These two primary inclusion criteria were intended to yield a minimally selected cohort of individuals in whom a clinician both diagnosed and treated MDD. For any individual patient, data were restricted to 2 years before and 4 years after initial AD prescription. In this restricted set of codes, we excluded patients with age below 18 at their first code and with age above 80 at their last code. To ensure sufficient coded data available prior to and following treatment, the cohort was limited to those whose first code (of any type) was observed at least 1 year before their first AD and whose last code was observed at least 2 years after their first AD. In all analyses, time was defined relative to the initial antidepressant prescription.

2.3. Modeling approach

2.3.1. Pruning step

We first sought to reduce dimensionality while enriching coded data to ensure that it was relevant to psychiatric phenotypes, recognizing that subsequent clustering would otherwise simply recapitulate patients' medical comorbidities. Conversely, beginning with a curated list of psychiatry-specific codes would limit the ability to discover depression subtypes that might be reflected in non-psychiatric codes – for example, in non-psychiatric comorbidity.

To select depression-relevant codes in a data-driven manner, we instead identified coded data correlated with simple psychiatric outcomes: binary-encoded variables that indicate the class of ADs prescribed to the patient (namely, selective serotonin reuptake inhibitor (SSRI), selective norepinephrine reuptake inhibitor (SNRI), MAOI, tricyclic antidepressant (TCA), or Other), whether the patient was prescribed more than three ADs in the two years following the initial AD prescription, or whether the patient's diagnosis was subsequently changed to bipolar disorder (for detailed patient outcome definitions, please refer to the S2 and S3).

We computed the Matthews correlation coefficient (Matthews, 1975) of each code with each outcome (a value between −1 and 1), and then all pruned codes that had the absolute value correlation below 0.01 for all outcomes. This threshold was chosen to be low so that even weakly correlated codes are used for the analysis. This step yielded 7121 codes, drawn from an initial 94,102 codes.

2.3.2. Representation learning step

To identify and categorize underlying disease concepts from diagnostic codes data, we trained a Latent Dirichlet Allocation (LDA) model (Blei et al., 2003). LDA is an unsupervised method for finding hidden thematic structures, called topics, within a collection of patient records with diagnostic code counts. It works by assuming each patient record is composed of a mixture of these topics, and each topic is characterized by a specific distribution of codes. This allows LDA to categorize patients based on the underlying disease concepts (i.e. topics). For a detailed list of top words per topic, please see S4. During the training phase of the LDA model, we summarized the total code counts into a 'bag-of-words' representation of the data for each patient and discovered 50 topics, corresponding to the underlying disease concepts (see Fig. 3). After the training phase, we divided each patient's trajectory into six equal time windows, each spanning one year; the first AD code fell within the second window. Using the trained LDA model, we inferred the topic distribution vectors for each time window and concatenated these distributions to form the features representing the patient's trajectory.

2.3.3. Clustering step

At this step, the patients' histories are summarized by their topic distributions over their histories. We then used agglomerative clustering with Ward's linkage criterion to cluster patients into 15 subpopulations. Agglomerative clustering is an unsupervised learning technique that iteratively merges clusters based on similarity, and Ward's linkage criterion calculates the distance between clusters by minimizing the total within-cluster variance. To evaluate stability of clusters qualitatively, we first compared topic representation between clusters derived from the two largest hospitals. We then tested each hospital-specific clustering against the original clusters using the adjusted Rand index (Steinley, 2004), a measure of similarity that ranges from −1 to 1, and the adjusted mutual information score, ranging from 0 to 1 (Vinh et al., 2009).

All analyses were conducted in Python, with LDA performed using the Gensim library (Řehůřek and Sojka, 2010), clustering and evaluation of clustering performed using the scikit-learn library (Pedregosa et al., 2011), and subsequent analysis done using standard computational libraries for Python (Harris et al., 2020; Hunter, 2007; Virtanen et al., 2020; Wes McKinney, 2010).

3. Results

The cohort included 423,636 patients who met the inclusion criteria based on MDD diagnosis. After the exclusion of 16,972 patients who lacked any RXNORM codes, 88,718 patients who lacked an AD treatment, 18,768 patients due to age restrictions, and 121,649 patients due to insufficient history 2 years before and 4 years after the first AD treatment, we were left with 136,371 patients (95,034 women [70 %]; 41,337 men [30 %]; mean [SD] age, 47.0 [14.0] years; see Fig. 1 for details).

We identified 50 code-based topics, reflecting data-driven groups of clinical presentations. These topics were then used to derive 15 clusters, manually annotated by one of the authors to simplify further discussion, on the basis of predominant topics as well as sociodemographic differences. (Of note, cluster numbers themselves are arbitrary and do not indicate any priority.)

Of the 15 clusters, 3 clusters (complex psychiatric comorbidity cluster (4), depression/anxiety cluster (6), and primary care depression/anxiety cluster (7)) were notable for the preponderance of depression and anxiety diagnoses in the outpatient setting. Among these, the complex psychiatric comorbidity cluster (4) stood out for its younger demographics (56 % were below the age of 40, as opposed to 34 % in the overall population) and elevated prevalence of attention-deficit/hyperactivity disorder (ADHD), group therapy, and psychotherapy, while the depression and anxiety cluster (6) was similar but reflected older patients (17 % and 28 % patients in age groups 18–30 and 40–50, compared to 32 % and 18 % for the complex psychiatric comorbidity cluster (4). The primary care depression/anxiety cluster (7) was similar to the depression/anxiety cluster (6) but reflected substantial medical comorbidity, primary care treatment of depression, and more male patients. (For formal contrasts, see Supplemental Materials and Methods).

Another group of clusters corresponded to patients who exhibited a higher prevalence of pain (pain and anxiety cluster (2) and surgery and pain cluster (5)). Among these, cluster the pain and anxiety cluster (2) were characterized by older patients on average, while the surgery and pain cluster (5) reflected more obesity and low back pain in particular.

A subset of clusters reflected predominantly obstetrics and gynecological (OB/GYN) diagnoses (pregnancy cluster (0), younger gyn cluster (1), younger gyn-primary care cluster (3)). Most notably, all patients in the pregnancy cluster (0) included at least 1 pregnancy code, while the younger gyn cluster (1) represented younger women with more gynecologic diagnoses and the younger gyn-primary care cluster (3) included more pain diagnoses. Besides the differences in gender and increased gynecologic diseases, patients in the younger gyn cluster (1) closely resembled the complex psychiatric comorbidity cluster (4) and the

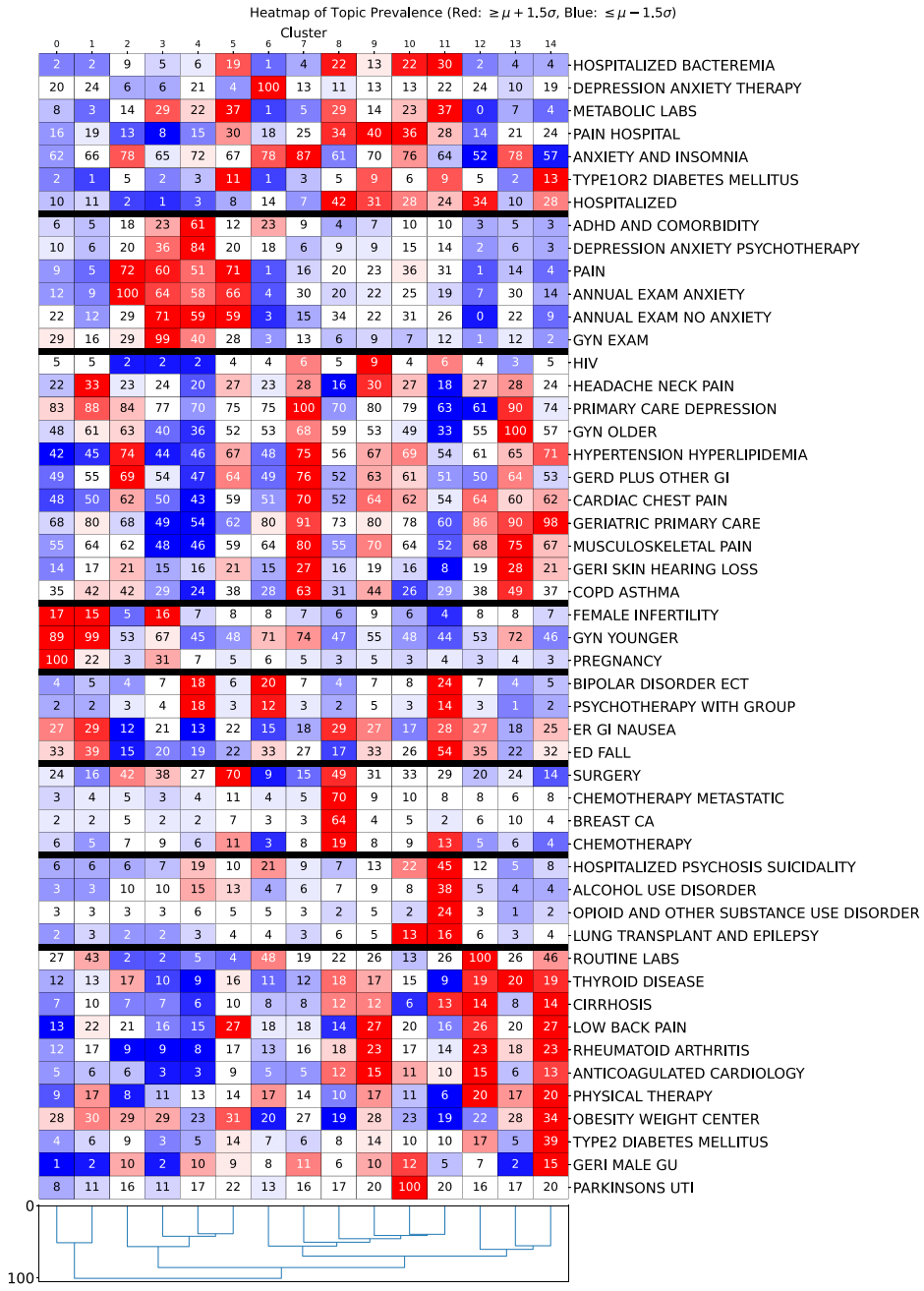

Fig. 3. Heatmap of topic prevalence.

depression/anxiety cluster (6), but with substantially greater rates of both substance use and suicidality. The older women cluster (13) was predominantly older women, with nearly all (93 %) 40 or older.

Finally, a group of clusters (breast cancer (8), general hospitalization (9), older men (10), hospitalization + hepatic disease (12), and obesity +DM2 (14)) captured individuals more likely to be hospitalized, varying by specific medical comorbidities (e.g., cancer in the breast cancer cluster (8) obesity and type 2 diabetes in the obesity and diabetes mellitus type 2 (DM2) cluster (14) and gender (e.g., nearly all males in the obesity and DM2 cluster (14)).

To understand the stability of clustering – i.e., the extent to which these clusters would be likely to generalize across hospitals and health systems – we compared results of clustering data from individual hospitals post hoc to those of the system as a whole. Supplemental Fig. S5 illustrates the representation of individual topics in each of the 15 clusters, derived from individual hospitals; with very few exceptions, the same topics drove cluster membership in each hospital. To formally test similarity of cluster assignments, we applied two alternate methods – the adjusted Rand index and adjusted mutual information. For comparison of MGH and BWH to system-wide clusters, adjusted Rand index was 0.31 for each; adjusted mutual information index was 0.46 for MGH and 0.44 for BWH. Thus, both qualitative and quantitative testing supported the lack of site-specificity of these clusters.

We then examined distribution of antidepressant treatments and outcomes across the clusters, as a way of understanding if clusters are meaningful in terms of understanding clinical differences. Figs. 4 and 5 illustrate the differences in each cluster for a given outcome, compared to the mean of the sample as a whole – i.e., the extent to which each cluster is different from the undifferentiated cohort in a particular outcome. The proportion of patient outcomes per group, color-coded by their deviation from the population average, can be found in S6. Chi-squared residuals and adjusted residuals for cluster-outcome

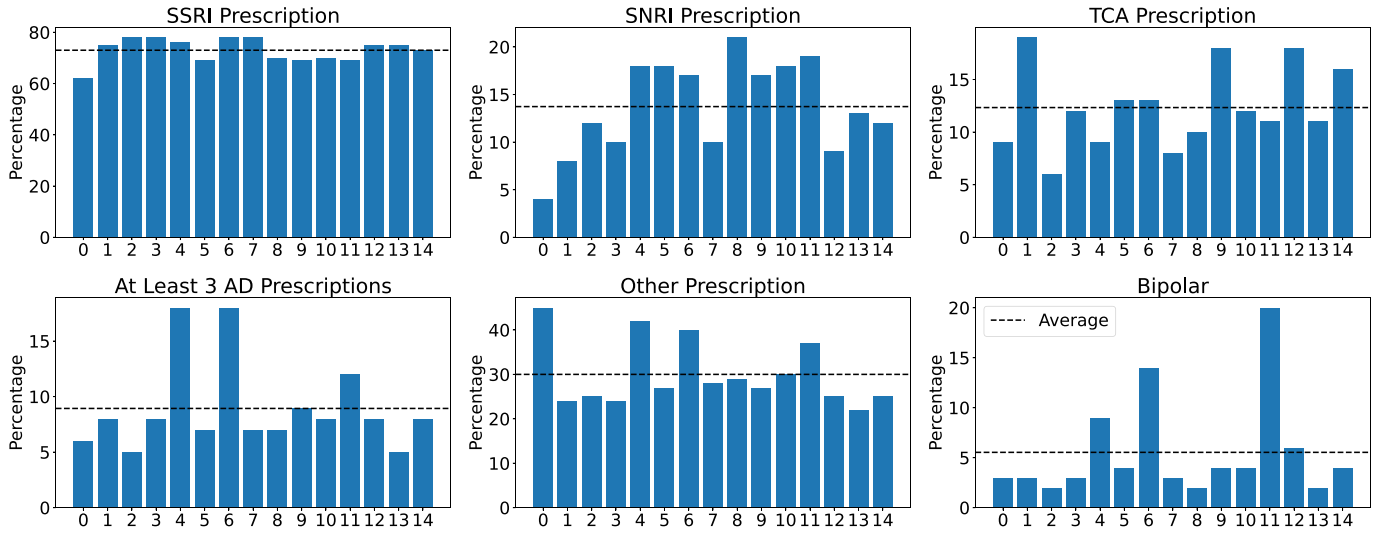

Fig. 4. Bar graphs of patient outcomes.

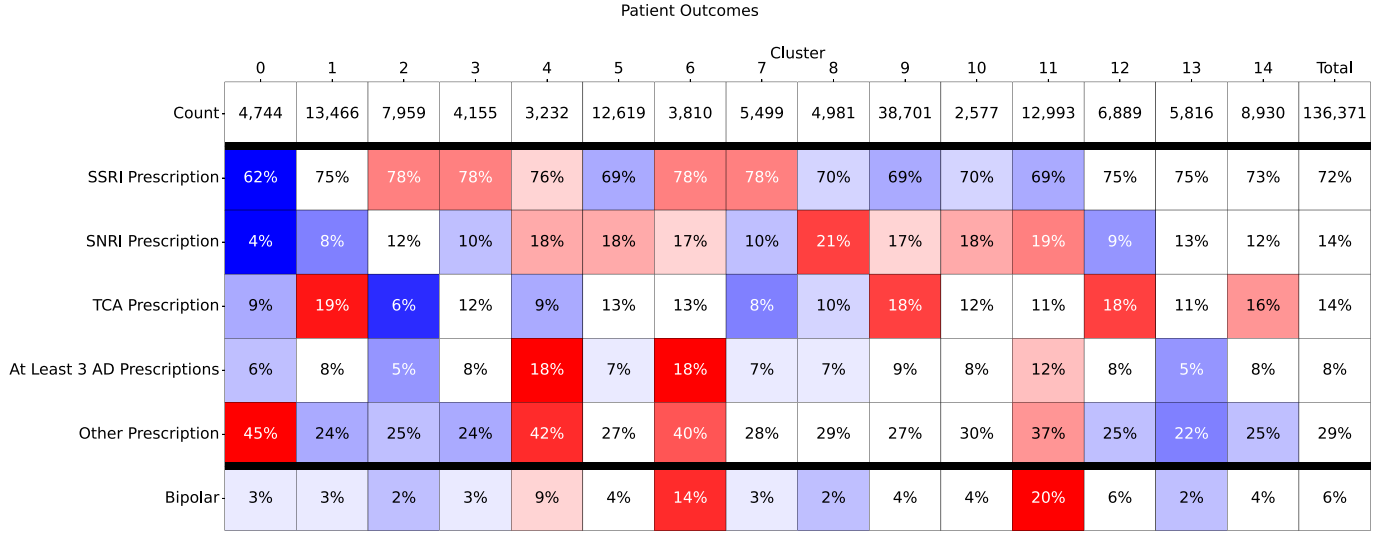

Fig. 5. Heatmap of patient outcomes.

relationships can be found in S7 and S8, indicating that nearly every cluster differs significantly from population mean values.

The complex psychiatric comorbidity cluster (4), depression and anxiety cluster (6), and the substance use and seizure cluster (11) were notable for markedly greater rates of change in diagnosis to bipolar disorder (9, 14, and 20 %, respectively), and use of 3 or more ADs in the 2 years following the first (18, 18, and 12 %, respectively). There was substantial variation in rates of SSRI use (from a low of 62 % in the pregnancy cluster (0), to 78 % in multiple other clusters) and SNRI use (4 % in the pregnancy cluster (0) and 8 % in the younger women/GYN group, up to 21 % in the breast cancer cluster (8)). Similar variability was apparent in other (non-SSRI/SNRI) AD use, from 22 % in the older women cluster (13) to 45 % in the pregnancy cluster (0). We identified 5 clusters with high levels of non-SSRI utilization. The breast cancer cluster (8) and the substance use and seizure cluster (11) reflected the highest rates of SNRI prescription and were notable for both primary care depression and geriatric primary care.

Finally, we inspected the change of topic prevalence over time for each of the clusters, as a means of understanding trajectories (see S9). Within the group of clusters described by complex psychiatric conditions, we observe that the complex psychiatric cluster (4) tended to exhibit greater prevalence of ADHD comorbidity over time, in contrast to the remaining clusters. The younger gyn cluster (1) was notable for its considerable increase in the prevalence of substance use and hospitalization after the index AD prescription. Among the OB/GYN clusters, the pregnancy cluster (0) showed increasing prevalence in head, neck, and lower back pain, while the younger gyn-primary care cluster (3) displayed an uptick in depression and anxiety following the initial AD prescription before returning to baseline.

4. Discussion

In this investigation of 176,529 adults with a diagnosis of MDD, we used unsupervised machine learning methods to generate 15 putative data-driven subtypes on the basis of coded data from the electronic health records of two hospital systems. Because we only clustered on codes that were associated with future psychiatric treatment or outcomes, we can interpret the resulting clusters as providing meaningful distinctions among patients with MDD. For example, the pregnancy cluster suggests that a person experiencing MDD around pregnancy is not an MDD patient who happens to be pregnant; the fact that they are pregnant affects correlates of their MDD.

Overall, the discovered subtypes differed markedly in terms of predominant sociodemographic features, comorbidities, and patterns of medication use. Beyond simply quantifying the variability within a single psychiatric diagnosis, each of these clusters identifies a more homogeneous, internally consistent group that may merit further characterization in terms of underlying neurobiology or treatment response. We are finding and separating constellations of medical comorbidities that matter in the context of MDD, rather than medical comorbidities that patients simply happen to have alongside their MDD.

This work extends longstanding efforts to parse the heterogeneity of major depression, employing a broad range of strategies that predominantly relied on expert-curated phenotypes. Some strategies simply employ age, distinguishing (for example) early onset depression (Zisook et al., 2007) or geriatric depression (Alexopoulos, 2019), where treatment response or tolerability may differ. Another typical approach examined individual depressive symptoms. A systematic review of studies that examined the existence of MDD subtypes by latent variable analysis of depressive symptoms found insufficient evidence for MDD data-driven symptomatic subtypes of depression (van Loo et al., 2012). However, the reliance in those efforts on symptoms did not allow for the possibility that systemic manifestations of illness, or even sociodemographic differences, may enable more precise characterization.

Yet another strategy has considered comorbid psychiatric or non-psychiatric illness, as potential markers of difference in underlying disease process. For example, one prior study computationally derived three depression subtypes that included patients who were the oldest, had the most comorbid diagnoses, and took the most medications (Xu et al., 2020) – essentially recovering geriatric depression. While we do identify significant variability in age distribution among the 15 clusters, our clusters are driven more by specific comorbidities than age per se.

This work has multiple strengths. First, it draws on a large and diverse outpatient population, served by multiple academic medical centers as well as community hospitals and their affiliate networks. That diversity should increase the generalizability of these results. By design, we included a broad age range and inclusion criteria, reasoning that this data-driven approach should be able to identify distinct populations rather than needing to specify them a priori (for example, by excluding geriatric depression). Furthermore, relying on large-scale health records rather than a single study or group of clinical trials should also diminish ascertainment bias, as might be the case in relying solely on individuals entering depression treatment studies. Finally, the application of unsupervised methods allows us to generate novel hypotheses about depression subtypes, rather than pursuing the same curated subtypes that have been a focus of research for many decades.

In aggregate, our study demonstrates the potential utility of unsupervised learning approaches applied to large-scale electronic health records to better understand the heterogeneity of MDD. The clusters identified by our modeling approach may facilitate efforts to characterize disease biology and develop predictive tools by incorporating the temporal aspect of patient history to reason about future outcomes. Further work will be needed to better understand the extent to which these clusters associate with differential course or treatment response, as well as their consistency in other regions of the US or internationally.

4.1. Limitations

We also note multiple limitations. Electronic health records lack reliable symptom-level data, so we cannot examine the extent to which subtypes might differ in clinical presentation or symptom dimensions (Kung et al., 2022). Our own prior work demonstrates that narrative clinical notes may in some cases provide such detail (Castro et al., 2014; McCoy et al., 2016; McCoy et al., 2015), although more recent narrative notes in Epic-based EHRs are highly templated and thus contain far more impoverished clinical descriptions. Moreover, NLP-based phenotyping transfers poorly across health systems, so incorporating such data would limit the ability of other health systems to replicate and extend our work. We also lack data on features such as social determinants of health, stressors, and social functioning more generally, features that might well associate with or even drive clusters.

A further limitation is the inclusion of only one regional health system, albeit one comprised of a heterogeneous group of academic and community hospitals and affiliate practices. While we show consistency of clusters across two very different large hospitals, further work will be required to understand the transferability of the putative subtypes we identified; it is possible that the clusters we observe are specific to features of care in this region. In addition, our reliance on antidepressant prescriptions to refine our definition of major depression may lead to a more select (albeit still quite broad) population, and the extent to which these clusters are also apparent for non-antidepressant-treated patients remains to be determined.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jad.2024.03.162.

Funding

This study was supported by the National Institute of Health (R01MH123804-03; Dr. Perlis and Dr. Doshi-Velez). The sponsors did not contribute to any aspect of study design, data collection, data analysis, or data interpretation. The authors had the final responsibility for the decision to submit for publication.

CRediT authorship contribution statement

Abhishek Sharma: Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Pilar F. Verhaak:** Writing – review & editing, Writing – original draft, Project administration. **Thomas H. McCoy:** Validation, Supervision, Investigation, Conceptualization. **Roy H. Perlis:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization. **Finale Doshi-Velez:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

Dr. Perlis has received consulting fees from Burrage Capital, Genomind, RIDVentures, and Takeda. He holds equity in Outermost Therapeutics and PsyTherapeutics. The other authors report no competing interests. Dr. Finale Doshi-Velez has no competing interested to report.

Data availability

Data from this manuscript are not available due to the limitations required by the Mass General Brigham human subjects review board regarding access to clinical data.

Acknowledgements

This study would like to acknowledge Sarah Rathnam and Isaac Lage for their assistance with data acquisition. The authors would like to thank Marton Havasi, Isaac Lage, and Sarah Rathnam for their assistance with pipeline implementation.

References

Alexopoulos, G.S., 2019. Mechanisms and treatment of late-life depression. Transl. Psychiatry 9, 188. https://doi.org/10.1038/s41398-019-0514-6.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Castro, V.M., McCoy, T.H., Cagan, A., Rosenfield, H.R., Murphy, S.N., Churchill, S.E., Kohane, I.S., Perlis, R.H., 2014. Stratification of risk for hospital admissions for injury related to fall: cohort study. BMJ 349, g5863. https://doi.org/10.1136/bmj.g5863.

Fang, Y., Fritsche, L.G., Mukherjee, B., Sen, S., Richmond-Rakerd, L.S., 2022. Polygenic liability to depression is associated with multiple medical conditions in the electronic health record: phenome-wide association study of 46,782 individuals. Biol. Psychiatry 92, 923–931. https://doi.org/10.1016/j.biopsych.2022.06.004.

Fava, M., Rush, A.J., Alpert, J.E., Carmin, C.N., Balasubramani, G.K., Wisniewski, S.R., Trivedi, M.H., Biggs, M.M., Shores-Wilson, K., 2006. What clinical and symptom features and comorbid disorders characterize outpatients with anxious major depressive disorder: a replication and extension. Can. J. Psychiatr. 51, 823–835. https://doi.org/10.1177/070674370605101304.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2.

Hart, K.L., Pellegrini, A.M., Forester, B.P., Berretta, S., Murphy, S.N., Perlis, R.H., McCoy, T.H., 2021. Distribution of agitation and related symptoms among hospitalized patients using a scalable natural language processing method. Gen. Hosp. Psychiatry 68, 46–51. https://doi.org/10.1016/j.genhosppsych.2020.11.003.

Hunter, J.D., 2007. Matplotlib: a 2D graphics environment. Computing in Science & Engineering 9, 90–95. https://doi.org/10.1109/MCSE.2007.55.

Hyman Rapaport, M., 2007. Translating the evidence on atypical depression into clinical practice. J. Clin. Psychiatry 68 (Suppl. 3), 31–36.

Jones, B.W., Taylor, W.D., Walsh, C.G., 2023. Sequential autoencoders for feature engineering and pretraining in major depressive disorder risk prediction. JAMIA Open 6, ooad086. https://doi.org/10.1093/jamiaopen/ooad086.

Kung, B., Chiang, M., Perera, G., Pritchard, M., Stewart, R., 2022. Unsupervised machine learning to identify depressive subtypes. Healthc Inform Res 28, 256–266. https://doi.org/10.4258/hir.2022.28.3.256.

Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) – Protein Structure 405, 442–451. https://doi.org/10.1016/0005-2795(75)90109-9.

McCoy, T.H., Castro, V.M., Cagan, A., Roberson, A.M., Kohane, I.S., Perlis, R.H., 2015. Sentiment measured in hospital discharge notes is associated with readmission and mortality risk: an electronic health record study. PLoS One 10, e0136341. https://doi.org/10.1371/journal.pone.0136341.

McCoy, T.H., Castro, V.M., Roberson, A.M., Snapper, L.A., Perlis, R.H., 2016. Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. JAMA Psychiatry 73, 1064–1071. https://doi.org/10.1001/jamapsychiatry.2016.2172.

Patel, R., Irving, J., Brinn, A., Taylor, M., Shetty, H., Pritchard, M., Stewart, R., Fusar-Poli, P., McGuire, P., 2022. Associations of presenting symptoms and subsequent adverse clinical outcomes in people with unipolar depression: a prospective natural language processing (NLP), transdiagnostic, network analysis of electronic health record (EHR) data. BMJ Open 12, e056541. https://doi.org/10.1136/bmjopen-2021-056541.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Perlis, R.H., Fraguas, R., Fava, M., Trivedi, M.H., Luther, J.F., Wisniewski, S.R., Rush, A. J., 2005. Prevalence and clinical correlates of irritability in major depressive disorder: a preliminary report from the sequenced treatment alternatives to relieve depression study. J Clin Psychiatry 66, 159–166 quiz 147, 273–274.

Řehůřek, R., Sojka, P., 2010. Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. ELRA, Valletta, Malta, pp. 45–50.

Rush, A.J., Fava, M., Wisniewski, S.R., Lavori, P.W., Trivedi, M.H., Sackeim, H.A., Thase, M.E., Nierenberg, A.A., Quitkin, F.M., Kashner, T.M., Kupfer, D.J., Rosenbaum, J.F., Alpert, J., Stewart, J.W., McGrath, P.J., Biggs, M.M., Shores-Wilson, K., Lebowitz, B.D., Ritz, L., Niederehe, G., STAR*D Investigators Group, 2004. Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. Control. Clin. Trials 25, 119–142. https://doi.org/10.1016/s0197-2456(03)00112-0.

Stavrakaki, C., Vargo, B., 1986. The relationship of anxiety and depression: a review of the literature. Br. J. Psychiatry 149, 7–16. https://doi.org/10.1192/bjp.149.1.7.

Steinley, D., 2004. Properties of the Hubert-Arable Adjusted Rand Index. Psychol. Methods 9, 386–396. https://doi.org/10.1037/1082-989X.9.3.386.

van Loo, H.M., de Jonge, P., Romeijn, J.-W., Kessler, R.C., Schoevers, R.A., 2012. Data-driven subtypes of major depressive disorder: a systematic review. BMC Med. 10, 156. https://doi.org/10.1186/1741-7015-10-156.

Vinh, N.X., Epps, J., Bailey, J., 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary?. In: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09. Association for Computing Machinery, New York, NY, USA, pp. 1073–1080. https://doi.org/10.1145/1553374.1553511.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, 2020. SciPy 1.0: fundamental algorithms for scientific computing in python. Nat. Methods 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

Wes McKinney, 2010. Data structures for statistical computing in Python. In: van der Walt, S., Millman, Jarrod (Eds.), Proceedings of the 9th Python in Science Conference, pp. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a.

Xu, Z., Wang, F., Adekkanattu, P., Bose, B., Vekaria, V., Brandt, P., Jiang, G., Kiefer, R.C., Luo, Y., Pacheco, J.A., Rasmussen, L.V., Xu, J., Alexopoulos, G., Pathak, J., 2020. Subphenotyping depression using machine learning and electronic health records. Learning Health Systems 4, e10241. https://doi.org/10.1002/lrh2.10241.

Zisook, S., Lesser, I., Stewart, J.W., Wisniewski, S.R., Balasubramani, G.K., Fava, M., Gilmer, W.S., Dresselhaus, T.R., Thase, M.E., Nierenberg, A.A., Trivedi, M.H., Rush, A.J., 2007. Effect of age at onset on the course of major depressive disorder. Am. J. Psychiatry 164, 1539–1546. https://doi.org/10.1176/appi.ajp.2007.06101757.