Brain Informatics

RESEARCH Open Access

Treatment journey clustering with a novel kernel k-means machine learning algorithm: a retrospective analysis of insurance claims in bipolar I disorder

Matthew Littman[1*], Huy-Binh Nguyen[1], Joanna Campbell[1] and Katelyn R. Keyloun[1]

Abstract

In real-world psychiatric practice, patients may experience complex treatment journeys, including various diagnoses and lines of therapy. Insurance claims databases could potentially provide insight into outcomes of psychiatric treatment processes, but the diversity of event sequences restricts analyses with currently available methods. Here, we developed a novel kernel k-means clustering algorithm for event sequences that can accommodate highly diverse event types and sequence lengths. The approach, Divisive Optimized Clustering using Kernel K-means for Event Sequences (DOCKKES), also leverages a novel performance metric, the transition score, which measures sequence coherence in individual clusters. The performance of DOCKKES was evaluated in the context of bipolar I disorder, which is characterized by heterogeneous treatment journeys. We conducted a retrospective, observational analysis of a large sample (n = 31,578) of patients with bipolar I disorder from the MarketScan® Commercial Database. Using insurance claims, bipolar episode diagnoses and mental health–related lines of therapy were identified as events of interest for patient clustering. The dataset included 202,122 events; 75% of the cohort experienced unique treatment journeys. Based on an optimal run, DOCKKES identified 16 treatment journey clusters, which were evenly split for initial manic/mixed or depressive episodes (8 clusters each) and varied in sequence length and early lines of therapy. Variability across clusters was also observed for demographics, comorbidities, and mental health–related healthcare resource utilization and cost. This proof-of-concept study demonstrated the use of DOCKKES for integrating information from large datasets, enabling comparisons between patient clusters and evaluation of real-world treatment journeys in the context of evidence-based guidelines.

Keywords Bipolar 1 disorder, Clustering algorithm, Machine learning, Real-world evidence, Sequence clustering, Treatment journey, Treatment pattern

1 Introduction

Patients with chronic psychiatric illness can experience a myriad of treatments during their lifetimes. Although evidence-based treatment guidelines exist, real-world prescribing may deviate from guidelines due to clinical judgment and factors excluded from randomized controlled trials, including comorbidities and concomitant medications [1–3]. Given these deviations, there is a need to identify patterns in longitudinal sequences of clinical events, referred to here as treatment journeys. However, associations between real-world psychiatric treatment journeys and outcomes remain underexplored due to the diversity of individual patient experiences. The complexities of treatment over time have not been satisfactorily captured by traditional analytical approaches,

*Correspondence:
Matthew Littman
matthew.littman@abbvie.com
[1] AbbVie, North Chicago, IL, USA

such as cohort descriptive analyses, linear/logistic risk score models [4], and Cox proportional hazards models [5]. This problem is better suited to data-driven machine learning approaches that are increasingly being leveraged for healthcare data [6].

Psychiatric treatment journeys have several challenges that must be considered when selecting data and methods for event pattern analyses. Treatment often occurs in both inpatient and outpatient settings, so claims data are appropriate to provide broad coverage across providers. The analytical method must be compatible with high sequence variability because of combination prescribing (polypharmacy) and long sequence length, for which variability compounds exponentially. Due to this complexity, a useful approach is to focus on event order rather than taking time-based approaches, such as temporal association rule mining [7]. Highly flexible methods like deep learning via graph neural network training could handle large, complex sequences [8] yet offer limited clinical insight into the rules governing any detected patterns. Process mining approaches, first developed by van der Aalst et al., offer an exciting means of investigating healthcare data for outcomes of treatment journeys [9]. However, process mining of large patient-level datasets must contend with each variant of a process [10]. When process variation is too great, the dataset must be broken down into smaller, more coherent clusters of processes. Therefore, to meaningfully mine highly variable real-world psychiatric processes, a suitable method is needed to cluster treatment journeys from large datasets.

Many methods of healthcare event sequence clustering have recently been proposed, with the general goal of grouping patients who experience similar clinical events in a similar order. Markov clustering is common, as in a two-step approach proposed for heart failure administrative records [11], yet it can be computationally restrictive. Rama et al. proposed a Temporal Needleman–Wunsch (TNW) algorithm for rheumatology electronic health records, which is robust regarding missing data yet lacks customizability in data representation [12]. Given the possible effects of guidelines on clinical practice, it would be interesting to cluster patients using a proposed generative adversarial network, which includes the assumption that clinical sequences are generated from a mixture of latent policies [13]. However, such a model would struggle to account for imperfections that can exist in claims data, as well as policy variations across regions and healthcare organizations. Furthermore, recent studies and review articles have proposed clustering to support sepsis phenotyping, asthma phenotyping, rheumatic and muscular diseases, and to identify sub-groups of patients with varying levels of cognitive decline, yet recent research has not focused on modelling clinical sequences with the variability and complexity of treatment journey data [14–17]. This may also be due to the unique challenges associated with clustering mental health data noted in a 2023 review article, with the outstanding call to action to develop more robust clustering techniques for clinical use in psychiatry [18].

Therefore, in order to enable future investigations (eg, process mining) in the mental health space, we describe in this study the development of a clustering technique uniquely adapted to psychiatric information available from insurance claims data. The algorithm was designed to accommodate sequences of uneven length with a variety of categorical clinical events. A kernel k-means approach was chosen for its substantial flexibility for event inputs. In this approach, a custom kernel function can be applied to non-linearly separable, categorical data to enable their clustering based on Euclidian distances [19, 20]. We also sought to develop a novel cluster quality metric that could interpretably measure sequence coherence within a single cluster. This algorithm sets the stage for future comparisons of real-world treatment journeys versus clinical guidelines, with the goal of identifying guideline divergence, associated factors, and patient outcomes.

Bipolar I disorder (BP-I) insurance claims were chosen as a use case for the new clustering algorithm to explore its ability to describe a disease with high heterogeneity among treatment journeys [21]. BP-I is a severe, chronic illness characterized by recurring episodes of manic, depressive, or mixed symptoms between periods of remission (euthymia) [22]. Several clinical course patterns, such as age of onset and rapid cycling, vary across patients and may correlate with clinical outcomes [21]. Pharmacotherapies for BP-I span many drug classes, including typical and atypical antipsychotics, mood stabilizers, antidepressants, benzodiazepines, and stimulants. Treatment choice may be influenced by pre-existing medications, comorbidities, and tolerability factors, often resulting in guideline-incongruent treatment plans [3, 22]. The variability in BP-I disease and treatment courses typifies the challenges of analyzing real-world treatment journeys in psychiatry.

2 Materials and methods

The analysis pipeline developed in this study is shown in Fig. 1. The pipeline began with preprocessing insurance claims to extract diagnostic/treatment events for patients with BP-I (Sect. 2.1), then patient event sequences were clustered with a novel approach of Divisive Optimized Clustering using Kernel K-means for Event Sequences (DOCKKES; Sect. 2.2). The main principles for algorithm development were to emphasize: cluster sizes by prioritizing larger clusters, cluster quality for clinical

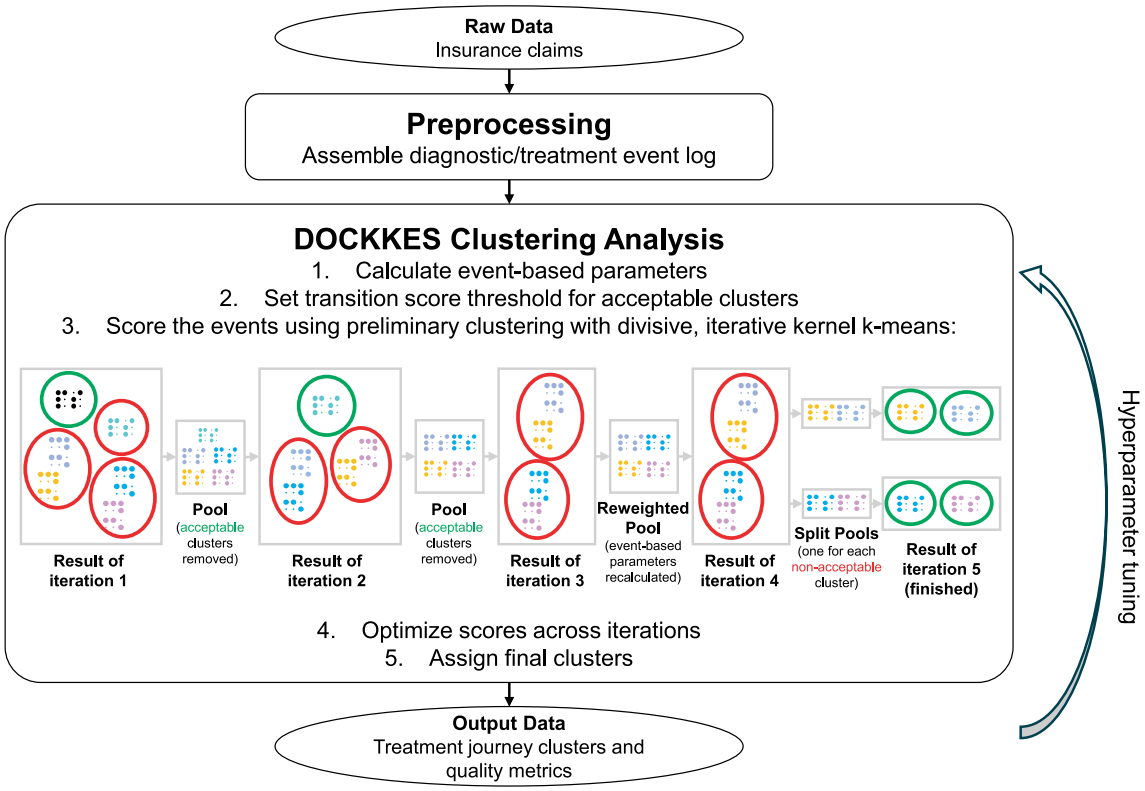

Fig. 1 Preprocessing and clustering pipeline. The DOCKKES schematic illustrates an example run of the divisive, iterative scoring procedure, with dots representing patients. Circles indicate acceptable (green) and non-acceptable (red) clusters, as determined by the cluster quality threshold. DOCKKES, Divisive Optimized Clustering using Kernel K-means for Event Sequences

interpretation by prioritizing similarity in treatment patterns within clusters, the reduction of complexity for transitions between events in resulting clusters, and weighting earlier events in a treatment sequence higher as earlier events may have greater clinical value versus later events. A key task in algorithm development was to identify a suitable kernel equation that captured event type and order, with appropriate weighting for clinically important factors such as common events and early sequence positions (Sect. 2.2.1). The temporal spacing of events was not factored into the kernel equation to limit problem complexity. Once the pipeline was developed, algorithm hyperparameters were tuned to determine an optimal set of treatment journey clusters (Sect. 2.3). Finally, the optimal clusters were analyzed descriptively, in which treatment journeys and additional patient-level clinical data were described among the patient clusters using Tableau (Tableau Software, Seattle, WA, USA) (Sect. 2.4). The pipeline (linked in the Availability of Data and Materials section) was implemented using Python custom-written scripts along with open-source packages, including Faiss, Hyperactive, Pathos, and PY-BOBYQA packages.

2.1 Description of dataset and preprocessing

2.1.1 Data source

This retrospective observational analysis used de-identified patient-level health data from the MarketScan® Commercial Database. The database included enrollment history and claims for medical and pharmacy services from United States employers and health plans representing covered employees and their dependents [23]. Because the database is primarily sourced from employers, it captures broad information on medication use by including claims for mail-order prescriptions and specialty pharmacies. Data complied with the Health Insurance Portability and Accountability Act.

2.1.2 Study population and design

The study design is shown in Fig. 2. The index date was defined as the first observed diagnosis of a BP-I manic, mixed, or depressive episode using International Classification of Diseases (ICD)-9 and ICD-10 diagnosis codes. Remission, bipolar II, and other bipolar disorder codes were not considered index diagnoses. The index diagnosis code could be in any position on an inpatient (hospitalization) claim or outpatient claim in any other

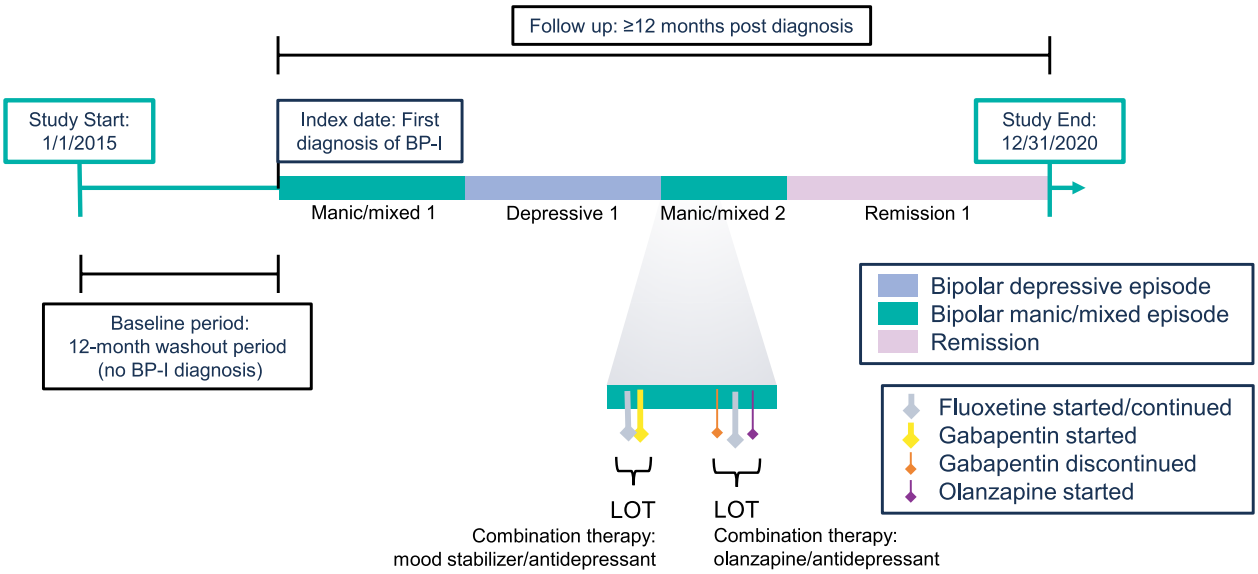

Fig. 2 Illustration of study design and BP-I event definitions. Events that were classified as inputs into the clustering algorithm were defined as the beginning of either a BPE or LOT. Example BPEs and LOTs are shown for illustrative purposes and do not represent an actual patient's events. BP-I, bipolar I disorder; BPE, bipolar episode; LOT, line of therapy

setting between January 1, 2016, and December 31, 2020. Patients were aged ≥ 18 years on the index date and were required to have ≥ 12 months of continuous enrollment before and after the index date. The baseline period was defined as the 12 months before the index date, and the follow-up period covered the index date to the earliest of the end of continuous enrollment or the end of study. Additional inclusion criteria were pharmacy, mental health services, and substance abuse coverage during the baseline and follow-up periods, as well as ≥ 1 line of oral therapy with a mental health–related treatment during the follow-up period. Patients were excluded if they had a prior BP-I diagnosis during the baseline period, a gap of > 30 days between continuously enrolled periods, or a claim for a long-acting antipsychotic injectable during baseline or follow-up (excluded due to inability to determine days of supply).

2.1.3 Preprocessing: assembly of event log

In order to group patients on the basis of the data most clinically relevant to their bipolar treatment journey, two events of interest were chosen as inputs to the clustering algorithm: bipolar episode (BPE) diagnoses and mental health–related lines of therapy (LOT; Fig. 2). Episode diagnoses, although not interventions, were included because of their importance in treatment selection. Both types of events were identified based on claims data available from the MarketScan Database, with some preprocessing to reduce treatment variability. Events for all patients were represented in a single event log. Each entry included (1) a patient identifier, (2) a BPE or LOT event identifier (described below), and (3) a timestamp for the event. The event log was filtered to remove event types that occurred for ≤ 0.05% of the patient population.

BPEs were identified from the bipolar ICD-9/ICD-10 diagnosis codes listed in Supplemental Table S1 and classified as manic/mixed (with or without psychosis), depressive (with or without psychosis), or remissive, as defined by the current or most recent episode listed in the code. BPEs were also numbered by position in each patient's sequence of episodes of that type (ie, a patient's first and second manic episodes were considered different event types: Manic Episode 1 and Manic Episode 2, respectively). The first episode for each patient began on their index date, and any subsequent episodes began on the date of an inpatient or outpatient medical claim of a different type from the prior episode.

LOT regimens within each BPE were identified based on pharmacy prescription fills or administrative procedure codes for on- or off-label mental health–related medications. Each LOT included all claims or codes within a 7-day window, plus medications continuing from prior periods based on the appended days of supply. To reduce treatment variability, medications were classified by drug generic name (all atypical antipsychotics) or drug class (typical antipsychotics, nonantipsychotic mood stabilizers, antidepressants, benzodiazepines, stimulants, and off-label treatments; Supplemental Tables S2–8). LOTs were reported by the list of all treatments and additionally classified as monotherapy (1 unique drug name/

class) or combination therapy (>1 drug name/class; eg, combination therapy: gabapentin/antidepressant). Start and end date definitions are detailed in Supplemental Table S9. Briefly, the start of the first LOT within each BPE was defined as the first claim or procedure code on or after the index/BPE start date. The start of subsequent LOTs within each BPE was defined by initiation of new medications beyond the 7-day window (including switches and augmentations) or reinitiation of a medication that was previously discontinued (defined as a gap in treatment of ≥45 days). Medications taken for ≤30 days were excluded from LOTs due to acute or trial use.

2.2 Description of the DOCKKES clustering approach

The DOCKKES approach encompasses three novel components that were developed specifically for the analysis of diagnostic/treatment event sequences of variable length: (1) a kernel equation ("event score formula"), (2) a new cluster quality metric for sequential categorical data ("transition score"), and (3) algorithm steps that leverage the event score formula and transition score to cluster patient event sequences.

2.2.1 DOCKKES event score formula

The event score formula acts as a kernel equation to represent non-linearly separable elements in a higher dimensional space. For DOCKKES, a formula was developed to represent each patient as a vector of scores for all event types in the entire dataset. If an event type happened more than once to a single patient, each instance of that event type was scored separately, then instances were summed. Event types that did not occur for a patient were scored as 0.

The event score formula was designed to capture positional information without allowing sequence length variability to dominate the analysis. It is defined for each instance of an event by Eq. 1:

$$score(event) = \frac{F_{event} + F_{next\ event}}{\left(P_{event}\left(a * R_{event,postion}\right)^{b} + P_{event}\left(a * R_{next\ event,next\ position}\right)^{b}\right)^{\frac{c}{clust\_num}}} \quad (1)$$

where F = total frequency of an event type among all patients;P = position of the event in the patient's sequence;R = rank by frequency of that event among all events that occurred in the same position (the most common event for that position is ranked at 1);*a, b,* and c = hyperparameters that the algorithm learns within user-inputted bounds; and *clust_num* = number of clusters assigned during the clustering iteration.

The F terms in the numerator of the event score formula give higher weight to current and subsequent events that occur more frequently in the dataset. The placement of P in the denominator gives less weight to events that occurred later in a patient's sequence based on the assumption that earlier events in a treatment journey were more predictive of a patient's overall path. The effect of P could be tuned by inputted bounds for the a and b hyperparameters (both restricted to ≥0). As a increased, the effect of an event's position was amplified b times.

Another goal in designing the event score formula was to limit the variability of the transitions between events to produce parsimonious clusters. This consideration was included by placing $R_{event,\ position}$ and $R_{next\ event,next\ position}$ in the denominator (eg, the event score was high if the patient went from the most common first event to the most common second event). This weighting allowed the clustering algorithm to recognize similar sequences by assigning equal values for the same transitions at the same position in an event sequence. This helped to reduce the number of unique transitions between events in a cluster. Both R terms are specifically amplified by the b hyperparameter so that DOCKKES can learn appropriate weights based on the dataset.

The c hyperparameter weighted the entire denominator, including P and R terms (c was also restricted to ≥0). As c increased, the clustering was primarily influenced by earlier events and more common event/position combinations. This was advantageous because there were fewer possible event types in the earlier positions (perhaps due to increasing LOT complexity over time), so higher weighting of earlier events could result in more distinct clusters. Since the placement of c in the exponent of the denominator would cause the algorithm to make c as high as possible, the exponent term was balanced with the number of clusters in the current iteration (*clust_num)* so that an increase in the number of clusters would reduce the impact of c. The placement of the *clust_num* variable favors fewer, larger clusters, depending on the inputted bounds for c. Although a high c term allows for a high number of clusters to reduce complexity and improve the interpretability of treatment journeys, too many clusters would not be manageable. Therefore, the upper bound of c is one of the most important decisions in tuning hyperparameters to a given dataset. Larger clusters are more valuable if the number of unique treatment patterns is low (which corresponds to a low maximum transition score as described in Sect. 2.2.2 below).

2.2.2 DOCKKES transition score

Common clustering quality assessment metrics, such as the silhouette average [24] and the Davies–Bouldin index [25], assess the quality of a set of clusters based on a combination of inter- and intra-cluster distance. Although these metrics explain cluster set quality in terms of spatial differencing between groups, they offer no insight into the quality of individual clusters in terms of sequence transition variability (ie, the number of unique transitions at each position in a sequence). Thus, we developed a new quality metric called the transition score, which was calculated per cluster by Eq. 2:

$$t(cluster) = \sum_{p=1}^{l} \frac{\#event\ types_p}{p} \tag{2}$$

where l = length of the longest sequence by any patient in that cluster; and p = position within the sequence.

High-quality clusters have smaller transition scores, indicating less diversity of events at each position in the sequence. Later positions contribute less to the transition score, supported by the assumption that earlier events in a patient's treatment journey are more clinically important. Therefore, minimizing the transition score in the DOCKKES algorithm incentivizes the similarity of event transitions across sequences, especially at earlier positions. The metric is used both during the DOCKKES algorithm to select acceptable clusters iteratively and to compare DOCKKES outputs and select for clustering runs with the lowest mean, maximum, minimum, or median transition scores across clusters.

In addition, the transition score has a theoretical maximum and minimum when given l and the number of event types because it is a finite summation series (specifically a harmonic series where the nth term approaches 0). As a result, the transition score of a single cluster can be evaluated relative to its theoretical minimum (eg, the closer to the theoretical minimum, the better the quality of the cluster). Consider the following example of a cluster, C, with three sequences of length ≤ 4 and a 4-event space (a, b, d, and e). Its sequences are (1) a > b > e > d; (2) a > b > d; and (3) b > a > e. The transition score for Cluster C would be calculated using Eq. 2 as follows:

$$t(C) = \frac{2}{1} + \frac{2}{2} + \frac{2}{3} + \frac{1}{4} = 3.917$$

This score can be compared with C's theoretical bounds for transition variability, calculated using Eq. 2 and assumptions of complete agreement of event types (best case) or complete disagreement of event types (worst case):

$$t(C)_{best} = \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} = 2.083$$

$$t(C)_{worst} = \frac{4}{1} + \frac{4}{2} + \frac{4}{3} + \frac{4}{4} = 8.333$$

which shows the example cluster is of decent quality, given its sequence lengths and event type diversity.

2.2.3 DOCKKES algorithm steps

The steps of DOCKKES are schematized in Fig. 1. The first three steps are presented as pseudocode in Table 1, and the remaining optimization steps are described below.

At the beginning of a DOCKKES run (step 1), the event log is used to calculate event-based parameters, which are later used to evaluate the event score formula.

Next, part 2 of DOCKKES sets a suitable transition score to use as a quality threshold for acceptable clusters identified during its iterative clustering procedure. The user can manually provide a transition score threshold or provide parameters to guide an automatic search for a threshold. This threshold serves as the quality gate for the next part of the algorithm. The number of clusters found in the threshold-setting run is stored as ν, a suggested number of clusters that will be used in step 5.

In the third part of the DOCKKES algorithm, events are iteratively scored, which accounts for the majority of the runtime. The goal was to create a meaningful numerical embedding of the patient's events and determine the number of clusters in the data. During each iteration, the transition score threshold could optionally be reduced by a set percentage each time. Reducing this limit helped if the sequences were very long and the initial transition score limit was primarily based on easy-to-find clusters. As the clusters are filtered out during each iteration, the quality threshold becomes more stringent, forcing the algorithm to search harder. The scoring events part ends when the maximum number of iterations has been reached or the algorithm has converged. Reweights are marked, event scores are collected for each person based on the optimized event score calculation of each iteration, and every clustering attempt is logged.

As shown at the end of Table 1, the initial DOCKKES scoring procedure in step 3 generates multiple scores (numerical embeddings) for the same patient and event type due to reweighting and splitting between iterations. In order to harmonize the iterations and determine one numerical representation for each event for each patient, step 4 of DOCKKES assigns a final weighting to each score column with a bounded non-distinct array of lambdas, or x. The objective is to learn the set of lambdas that maximizes Eq. 3:

Table 1 Pseudocode for key DOCKKES steps

Step 1: Calculating event-based parameters	**Input**: Event log, bounds for *a*, *b*, and *c* hyperparameters **For** events in event log Calculate inputs into event score formula **Return**: P_{event}, $R_{event,position}$, $R_{next\ event,next\ position}$, F_{event}, $F_{next\ event}$
Step 2: Setting the transition score threshold	**New Input:** Cluster number range, method of calculating threshold based on cluster transition scores (mean [default], max, min, median, std, x*[max–min]) **For** number of clusters in range: Run k-means algorithm using kernel event score formula and optimization of *a*, *b*, and *c* hyperparameters Find best quality run (based on Silhouette avg/Davies–Bouldin index + 1) Compute threshold using inputted method for the best quality run **Return:** Transition score threshold, number of clusters from best run (*v*)
Step 3: Scoring the events	**New Input:** Max number of iterations, cluster number range, min allowable cluster size, optional: % reduction in threshold for each iteration Set *reweighted* = 0 **While** iteration ≤ max Set *stuck* = 1 Reduce transition score threshold (optional) **For** number of clusters in range: Run k-means algorithm using kernel event score formula and optimization of *a*, *b*, and *c* hyperparameters Find best quality run (based on lowest max transition score) **For** clusters in best run **If** cluster transition score < threshold AND cluster size > min allowable Remove patients in cluster from pool Set *stuck* = 0 **If** all patients have been removed, **then break** **If** not *stuck*, **then** iteration + = 1, **continue** **If** not *reweighted* Recalculate P_{event}, $R_{event,position}$, $R_{next\ event,next\ position}$, F_{event}, $F_{next\ event}$ for remaining events Set *reweighted* = 1 iteration + = 1, **continue** **Else** Split: Initiate parallel iteration with each non-acceptable cluster Set *reweighted* = 0 iteration + = 1, **continue** **Return:** Matrix of scores by event type, patient, and iteration

$$f(x) = argmin\left(\sum_{c=0}^{g(x)} \frac{n_c}{t_c}\right) \quad (3)$$

where

$x = (\lambda_{i=1}, \ldots \lambda_k) \in \{0, \ldots, 5\} with\ a\ 0 \leq step\ size < 1$
$where\ k * len(list(np.arange(0{,}5, step\ size))) < 120$

- o k = number of score columns from step 3 of DOCKKES

$g(x) = Cluster(\sum_{i=1}^{k} \lambda_i s_i)$ is the result of clustering the weighted score columns

- o λ_i = Lagrangian coefficient to find for each score column s
- o s_i = each score column

n_c = number of patients in Cluster c
t_c = transition score for Cluster c

The bounds of 120 and 5 were selected to limit lambda search space possibilities so that memory overflow does not occur and no single event score column dominates the overall score. The $\frac{n_c}{t_c}$ term encourages fewer (and therefore larger) clusters with n in the numerator and lower transition scores with t in the denominator. Maximizing the clustering run with the lowest $\frac{n_c}{t_c}$ ensures that other clusters must have higher values, which results in all clusters being of good quality. DOCKKES uses two optimizers in parallel to learn the optimal set of λ s (Bayesian optimizer and the Repulsing Hill Climbing optimizer, both from the Hyperactive package). This process (step 4) results in a single optimized numerical embedding of each event for each patient.

Finally, step 5 uses the final event numerical embeddings from step 4, along with part 2's suggested number of clusters, v, to assign cluster labels to each patient. A final cluster search space is built with v as the minimum number of clusters and $2v$ as the maximum. For each

number of clusters in this range, a kernel k-means clustering run is performed. In order to balance the goal of larger clusters, *lc*, and simpler clusters, *sc*, the best number of clusters is chosen based on the clustering result that has the lowest value for $sc - lc$,

where $lc = mean(\frac{n_c}{t_c})$ across all clusters within the run, with min/max normalization of all *lc* values in the *v* to *2v* search space; and $sc = median(t_c)$ across all clusters within the run, with the same min/max normalization.

The median and mean were chosen for *sc* and *lc*, respectively, based on the potential for outliers, although using the same statistic for both would likely not affect the results. Because of the normalization, both *sc* and *lc* have values ranging from 0–1. Therefore, the best clustering run is one in which *sc* is minimized and *lc* is maximized. This approach addresses a challenge from the previous step, which is that $\frac{n_c}{t_c}$ improves (gets larger) with fewer clusters since each cluster will have more patients, but each transition score (t_c) improves (gets smaller) with more clusters since the complexity of each cluster can decrease. On the basis of the best clustering result according to this optimization, cluster labels are then assigned to each patient. The cluster label number represents group membership and does not indicate any ordering scheme.

2.2.4 DOCKKES efficiency

Algorithm speed and the ability to deal with large data sizes were crucial factors guiding the development of DOCKKES. The DOCKKES event score formula and algorithm were designed to reduce a 3-dimensional problem (patients, event types, and event position) to a 2-dimensional problem (patients and events, including type and position information), which substantially improves the speed of optimization versus existing clustering algorithms. While this may reduce the value of events that are more frequent, the impact does not outweigh the benefit to speed, as DOCKKES separately weights higher frequency events as described in Sect. 2.2.1. Additionally, the choice of kernel k-means clustering is advantageous as it allows the algorithm to run in linear dimensions of time and space, $O(n)$, in contrast to alternatives like hierarchical clustering or TNW which runs in higher complexities of time and space, $O(n^2)$ or even agglomerative hierarchical clustering which runs in $O(n^3)$, resulting in computational and memory efficiency [26]. For example, compared to agglomerative hierarchical clustering, the estimated time to cluster 1 million patients would be 31 years, whereas for DOCKKES the estimated time is 2.7 h. Particularly for larger psychiatric cohorts, algorithms must be capable of handling larger sample sizes; Indeed, the algorithm mentioned in Rama et al. [12] had 426 patients, but failed to complete on this study's larger BP-I cohort due to memory constraints. Other important advantages for applying DOCKKES versus existing clustering algorithms to psychiatry data is that due to large treatment variability, DOCKKES includes the ability to select minimum cluster sizes to ensure clinically meaningful clusters. DOCKKES is uniquely suited for treatment patterns data and makes a key assumption that earlier treatments are weighted higher than later events; yet, this assumption may not apply to all clinical psychiatric populations or non-psychiatric data. Parallel processing was also used whenever possible for initial cluster range searches, optimizations, and transition limit calculations.

DOCKKES allows the choice of optimization method from two open-source Python packages: Hyperactive (with parallel processing via Pathos package) and PYBOBYQA. When Hyperactive is selected, the DOCKKES algorithm chooses the best performer from its two parallel-run optimizers (Bayesian and Repulsing Hill Climber). Both optimization methods available are "derivative-free"—that is, they do not require derivative information to find local optima, and their use can reduce inaccuracy/noise and expense [27]. When the function to minimize has no clear derivative, derivative-free optimizers can still perform well to provide a theoretical convergence [27]. This study used the Hyperactive option.

2.3 Hyperparameter tuning and selecting an optimal clustering run

As with any machine learning algorithm, turning hyperparameters in DOCKKES is optional, but can considerably improve the output results. In this study, varying hyperparameter inputs were explored based on their impact on resulting clusters and the time necessary to run iterations; a total of 365 clustering runs were attempted. The parameters identified as having the largest impact on the resulting clusters are shown in Table 2.

DOCKKES uses the input ranges of parameters to find the optimal parameter. The numbers chosen as hyperparameters in this study are a result of many runs to find the boundaries of effect. For the cluster number range, higher ranges might mean that small clusters in the data may be removed from the iteration. The maximum number of iterations may allow the algorithm to run more iterations, yet may increase run-time. For this study, when c exceeds 2, it dominates the event score formula and reduces utility of the other hyperparameters. The 'a' hyperparameter was most impactful for this study in small positive integers, but can always be increased as long as it remains positive. The cluster number range for threshold setting should be a range that includes an estimate into how many clusters the

user believes may exist within the data. Each hyperparameter increases the total number of runs when searching across many hyperparameter sets.

Once hyperparameter tuning runs were completed, an optimal run was selected using a combination of quality metrics before further clinical analysis. Each clustering run was assessed by the Davies–Bouldin index, silhouette average, and mean transition score across its clusters. The three metrics were min/max normalized across all runs; silhouette averages were normalized in the opposite direction so that lower scores were better for each normalized metric. A total ranking was assigned by summing each of the three normalized values with equal weighting and ordering runs from best to worst. The optimal run by this combined metric was used for all clinical analyses in this study.

2.4 Analysis of optimal clustering run

The optimal run was assessed for BPE and LOT sequence trends. Other variables of clinical interest, demographics, and baseline characteristics were collected from the MarketScan Database and descriptively assessed across treatment journey clusters. Age, sex, region, and health plan type were assessed at each patient's index date. Mental health–related comorbidities (Supplemental Table S10) and Charlson comorbidity index (CCI) were assessed during the baseline period. Healthcare resource utilization (HRU) and cost data were also collected based on each patient's follow-up period (calculations described in Supplemental Table S11). Unadjusted, descriptive trends in variables of clinical interest, demographics, and baseline characteristics across clusters were visualized using Tableau.

3 Results

After applying inclusion and exclusion criteria, a total of 31,578 BP-I patients were included in the cohort (Fig. 3). The patients experienced a total of 202,122 BPE and LOT events, with 75% of the cohort (23,525) experiencing unique treatment journeys. The average duration of follow-up was 2.36 years, with an average of 1.72 BPEs and 4.68 LOTs per patient. The original number of event types (7671) was filtered down to 226 event types during preprocessing, which removed events that occurred at low frequency. Additional cohort demographics are shown in Supplemental Figure S1.

By varying hyperparameter inputs, 365 different DOCKKES clustering runs were conducted based on the patients' BPE and LOT event sequences. Each run was assessed with the novel transition score, as well as two standard clustering performance metrics (Davies–Bouldin index and silhouette average). On average, the 365 clustering runs had adequate performance across metrics (Table 3). In order to ensure correspondence between the performance metrics, each was normalized and plotted for all runs (Fig. 4). Each metric showed a similar distribution across runs, indicating the metrics generally correlated with one another.

The normalized performance metrics were summed with equal weight to identify the optimal clustering run, which yielded 16 clusters. The optimal clustering run had a mean transition score of 161.64 (Fig. 4 and Table 4). This run did not have the overall minimum mean transition score (154.26), suggesting that alternative weighting schemes for choosing the optimal clustering run could also be valid. The 16 clusters from the optimal run ranged in size from 259 to 9,554 patients (Fig. 5).

Clinical face validation of DOCKKES treatment journey clustering was assessed based on BPE and LOT trends across clusters from the optimal run. The resulting clusters captured variability of index BP-I diagnosis, with

Table 2 Impactful hyperparameter input variations

Parameter	Definition	Hyperparameters used
Cluster number range	An initial search into how many clusters there are likely to be	2–8, 3–8, 4–8, 4–9, 4–20
Maximum number of iterations	The maximum number of iteration steps to perform while running the algorithm	20, 25, 30
Maximum exponent (upper bound for the *c* hyperparameter)	A larger max exponent means that events that occur later in the sequence will contribute less to the score and, therefore, the clustering will focus more on earlier events	1.5, 1.6, 1.7, 1.8, 1.9, 2
Maximum slope (upper bound for the *a* hyperparameter)	Amplifies the ranked occurrence (*R*) of both the event and the next event; this parameter may put more weight on one or both events, depending on the balance between *R* for the current and next event	1, 2, 3, 4
Cluster number range for threshold-setting step	The range of clusters used to calculate the threshold that determines whether a cluster is of high enough quality or not; this is one of the first steps in the algorithm	7–14, 14–21

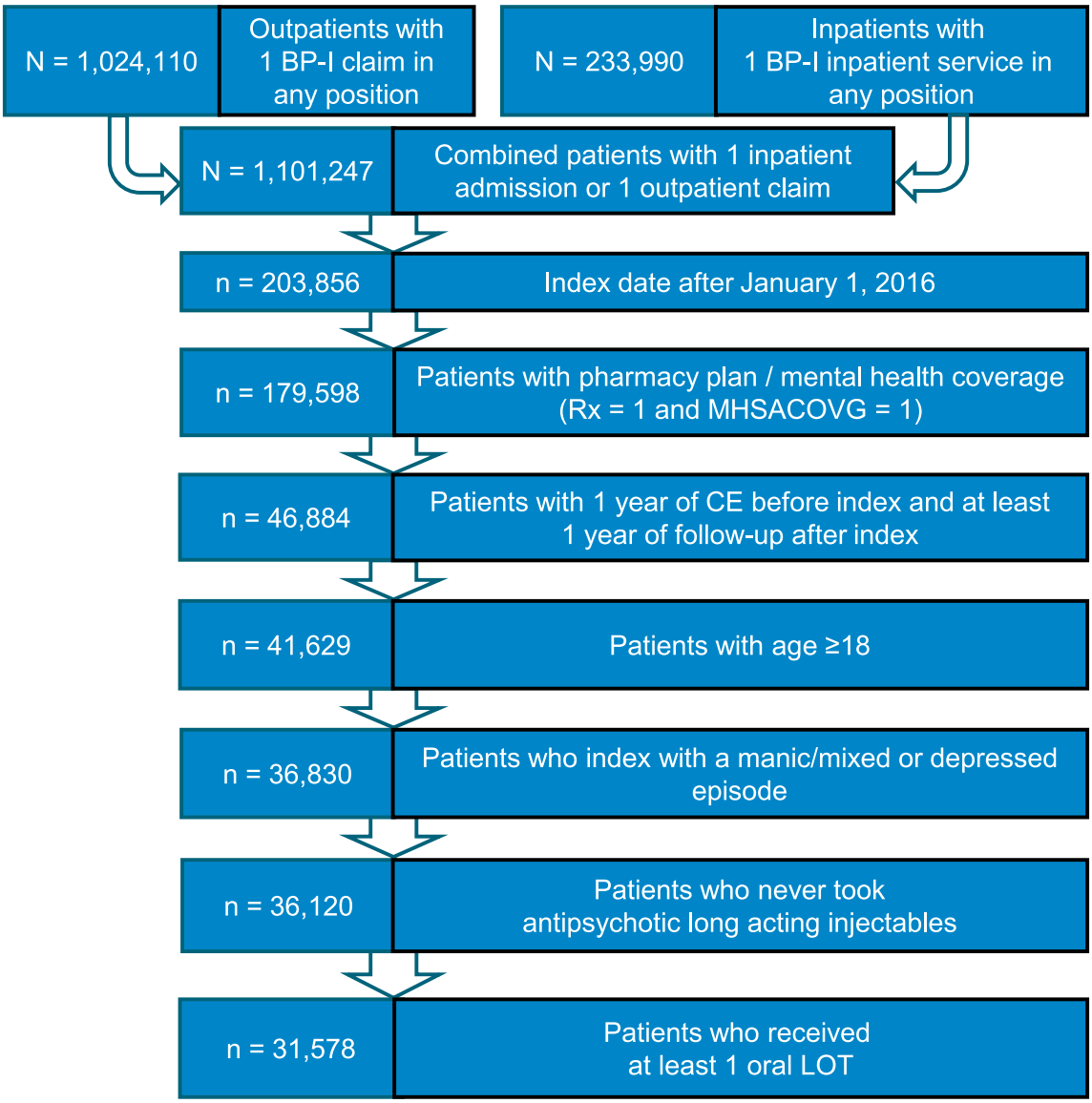

Fig. 3 Study attrition. BP-I, bipolar I disorder; CE, continuous enrollment; LOT, line of therapy; MHSACOVG, mental health services and substance abuse coverage; Rx, prescription/pharmacy plan

Table 3 Summary of all 365 clustering runs

	Reference range	Best run	Worst run	Mean of runs	Median of runs
Davies–Bouldin index	0 to infinity [0 is best]	0.319	2.558	0.919	0.817
Silhouette average	− 1 to 1 [1 is best]	0.897	0.388	0.723	0.738
Transition score (mean of clusters)	Depends on sequence length and number of event types [lower is better]	154.26	218.32	180.28	176.52

all patients in 50% (8/16) of the clusters exhibiting an initial manic/mixed episode (n = 15,355 total patients from clusters 0, 2, 4, 9, 10, 11, 12, and 13). In the remaining eight clusters, all patients exhibited an initial depressed episode (n = 16,223 total from clusters 1, 3, 5, 6, 7, 8, 14, and 15). The average number of episodes per patient also differed across clusters, with one treatment journey cluster (13; n = 547) having both the most manic/mixed

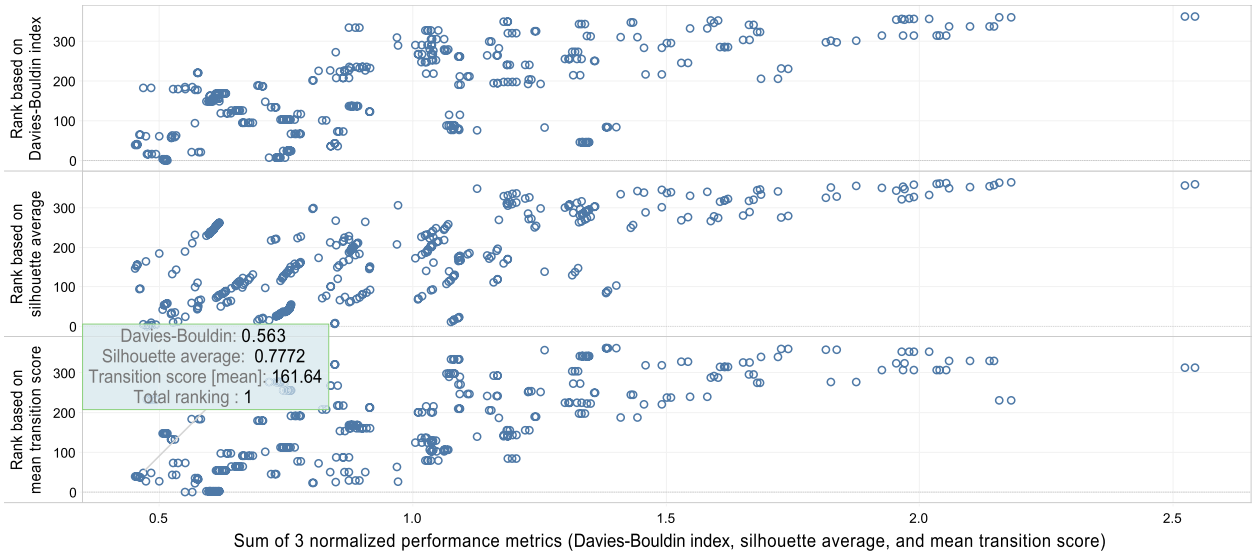

Fig. 4 Comparison of novel transition score and standard performance metrics. Dots represent 365 independent DOCKKES clustering runs with varying hyperparameter inputs. The normalized performance metrics each ranged from 0 (best score) to 1 (worst score) and then were summed to yield x-axis values between 0 and 3. Non-normalized metrics for the optimal clustering run are highlighted by the blue box

episodes (1.65 on average), the most depressed episodes (1.33 on average), and, consequently, a substantially greater number of total episodes (3.14 on average; Fig. 6).

Variability of LOTs was also observed across treatment journey clusters (Supplemental Fig. S2). Antidepressants and lamotrigine monotherapy were the most common regimens out of the first 3 LOTs overall; together, they were the first LOT for most participants in 9/16 clusters. Benzodiazepines were less common overall (prescribed for 3609 [11.4%] of all patients) but were the first LOT for a strong majority of patients in clusters 8 and 9. For the remaining clusters (0, 2, 5, 7, and 13), there was high variability in the first LOT, and no initial treatment accounted for > 20% of patients in a given cluster.

Demographic characteristics differed numerically across treatment journey clusters (Fig. 7). There were numerical differences in age, with almost a 10-year difference in average age between clusters 9 and 13. There were also inter-cluster differences in sex (21% range in the proportion of females between clusters 2 and 15: 49% and 28%, respectively) and CCI of 2 or more (range: 1.0% to 5.0% of patients in clusters 3 and 12, respectively).

In addition, treatment journey clusters demonstrated numerical differences in psychiatric comorbidity prevalence, which may have been related to treatment patterns (Fig. 8). For example, attention-deficit/hyperactivity disorder (ADHD) was more prevalent in clusters 0, 5, and 7 than other clusters. In two of these (clusters 0 and 7), stimulants were prevalent in the first 3 LOTs relative to most other clusters (Supplemental Fig. S2). However, several clusters with lower ADHD prevalence had relatively high stimulant prevalence (eg, cluster 11), suggesting non-ADHD–related stimulant use.

Cost and HRU numerical differences were observed across clusters (Fig. 9). Average BP-I–related cost diverged by more than $6,000 when comparing the

Table 4 Performance of the optimal clustering run

	Total of all clusters	Minimum cluster	Maximum cluster	Median of clusters	Mean (SD) of clusters
Unique treatment patterns	23,525	39	8524	518	1470.31 (2715.11)
Patients per cluster	31,578	259	9554	855	1973.62 (2982.27)
Transition score	2586.31	12.43	518.71	156.99	161.64 (162.61)
Silhouette average (– 1 to 1)	0.777				
Davies–Bouldin (0 to infinity)	0.563				
Number of clusters	16				
Runtime elapsed	34 min 50 s				

SD, standard deviation

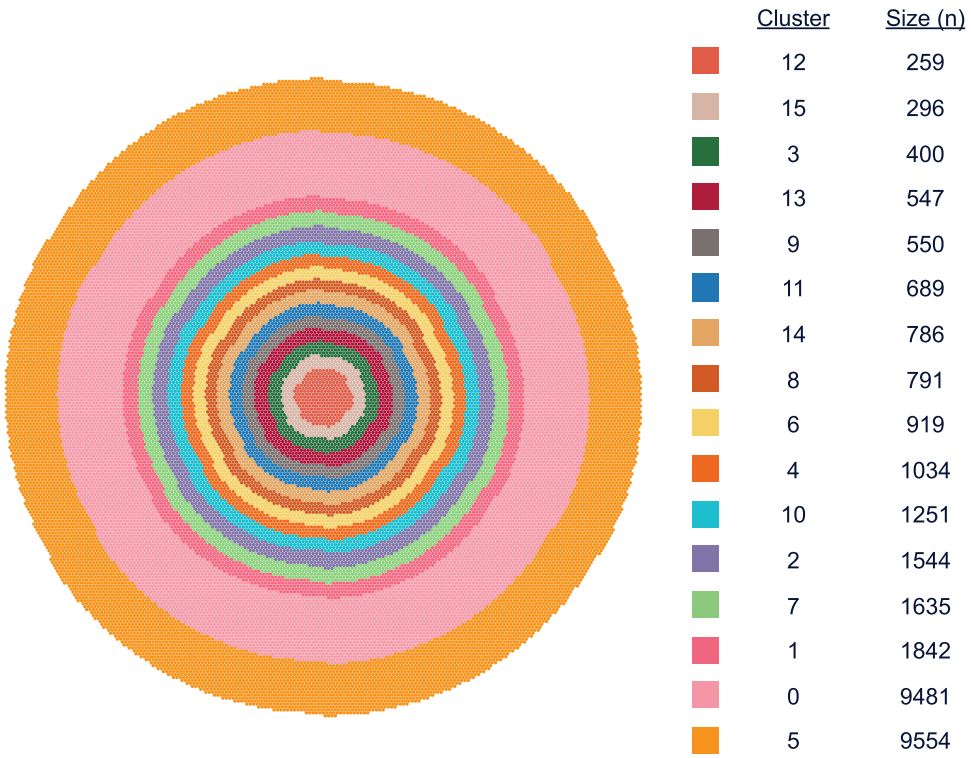

Fig. 5 Number of patients per cluster. Each color ring represents a cluster, with each dot representing a single patient. Clusters are ordered from the smallest cluster in the center to the largest on the outer rim

highest and lowest average costs (clusters 13 and 3, respectively). The same clusters differed by almost $10,000 in average mental health–related cost. Substantial ranges were also observed in average mental health–related HRU categories, including the number of pharmacy prescriptions (22.3 to 5.3 average annual prescriptions), outpatient services (13.0 to 6.8 average annual visits), and inpatient visits (0.9 to 0.6 average annual visits). Similarly, BP-I-related HRU differed when comparing outpatient services (average of 6.9 visits in cluster 13 and 2.4 visits in cluster 11) and inpatient services (average of 0.73 visits in cluster 13 and 0.50 visits in cluster 12).

Due to DOCKKES's sequence-based approach, additional clinical insight can be gained by investigating the order of events within each cluster. For example, a direct comparison of clusters 2 and 13 revealed that despite all patients in both clusters having the same index BPE (manic/mixed), clinically meaningful divergence occurred in the events directly following the index episode (Fig. 10). In cluster 2, observed second events were either a mental health–related treatment or remission (5% of cases), whereas in cluster 13, all patients experienced a depressive episode as their second event without receiving treatment for their index manic episode. Once they received treatment, patients in cluster 13 exhibited the third highest variability in their first LOT (Supplemental Fig. S2). Annual average mental health– and BP-1–related costs for cluster 13 were twice that of cluster 2 (Fig. 9). Affective switching between mania and depression is a hallmark of bipolar disease [22], thus validating that clusters identify clinically relevant and potentially differentiated patient populations.

4 Discussion

This proof-of-concept analysis using DOCKKES, a novel approach for treatment journey clustering, analyzed BP-I patients from a large US insurance claims database. We identified 16 unique treatment journey clusters from follow-up periods of over 2 years and over 6 events per patient on average. The resulting clusters supported the internal validity of the algorithm based on a weighted average using the novel transition score and standard clustering algorithm performance metrics. Clusters were balanced for initial depressive and manic episodes (8 clusters each), with approximately 50% of patients represented in either case. Substantial variation was observed among clusters in HRU and cost outcomes, as well as baseline demographics (eg, age and sex) and clinical characteristics (eg, CCI and

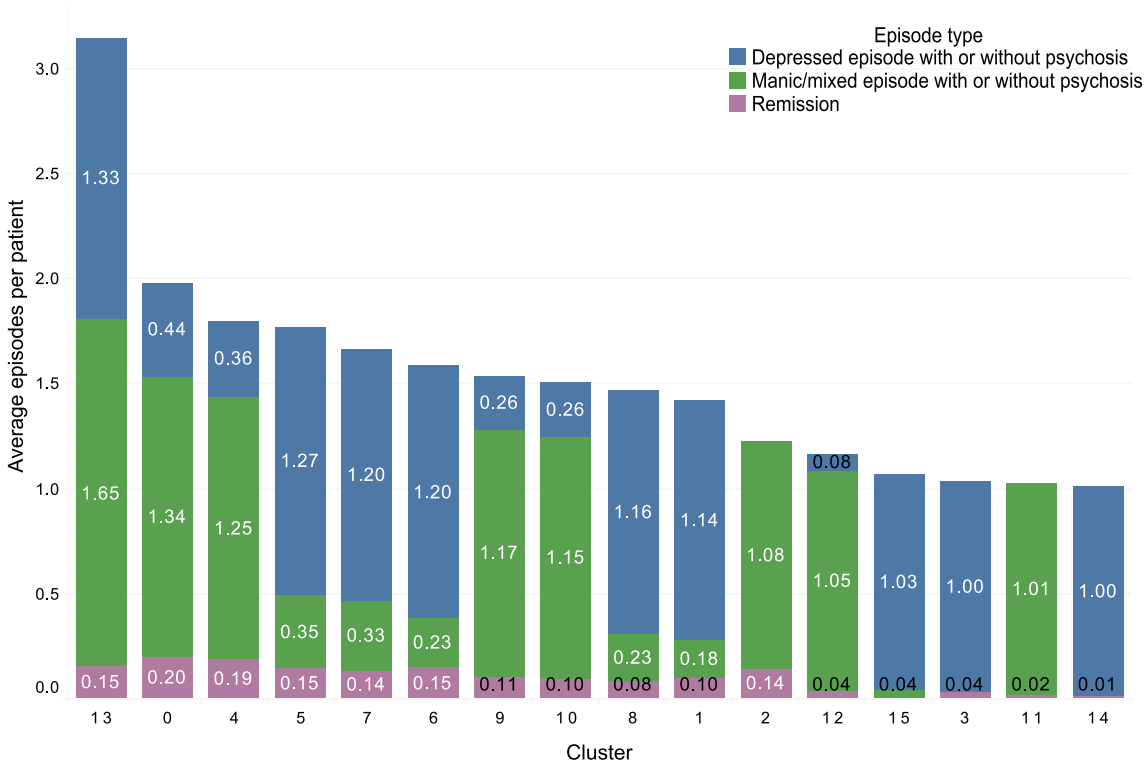

Fig. 6 Average number of episodes per cluster stacked by episode type. In each cluster, episode types with < 0.01 average episodes per patient were filtered out

mental health comorbidities). Given the large sample size, the degree of variation suggests that the algorithm may be clustering patients in a clinically meaningful way.

The identified treatment journey clusters can be evaluated in the context of BP-I treatment guidelines to assess preliminarily whether patients received guideline-congruent treatment. For example, patients in Cluster 3 experienced an initial depressive episode; the majority were treated with the mood stabilizer lamotrigine, which is one recommended first-line therapy option for bipolar depression [22]. Cluster 3 patients did not experience another episode on average during follow-up. In contrast, Cluster 13 patients did not receive pharmacologic treatment between their index manic episode and a subsequent depressive episode. This suggests that their treatment was not aligned with the recommendation to initiate a front-line medication for any case of acute bipolar mania [22], although the timing between the mania and depression events may need to be investigated. Cluster 13 patients also had the third highest variability in their first LOT, once initiated, and the highest average number of BPEs. Interestingly, Cluster 13 also had the overall highest mental health–related inpatient HRU, outpatient HRU, and total cost, whereas Cluster 3 had the overall lowest average cost and substantially lower HRU than Cluster 13.

Several other treatment journey clusters were identified that may have incongruency with BP-I guidelines based on specific drug classes. Benzodiazepines are not recommended as first-line therapy, although some guidelines suggest their use as adjunctive treatment for acute mania [3, 28]. However, benzodiazepine monotherapy was prescribed as the first LOT for most patients in clusters 8 and 9. Stimulants are generally not recommended except in cases of comorbid ADHD [22]. Although stimulant treatment correlated with ADHD comorbidity in some clusters, cluster 11 was a notable exception with relatively high stimulant treatment, low ADHD prevalence, and mostly manic/mixed episodes. These observations warrant investigation into the role of comorbidities and other factors in driving treatment patterns that involve potentially guideline-incongruent drug classes.

This study demonstrated the feasibility of clustering psychiatric patient treatment journeys using DOCKKES, which is capable of analyzing large-scale datasets (n = 31,578 patients) due to its modification of the efficient, commonly used kernel k-means algorithm. Additional advantages of the DOCKKES approach include flexibility for many treatment types and variable

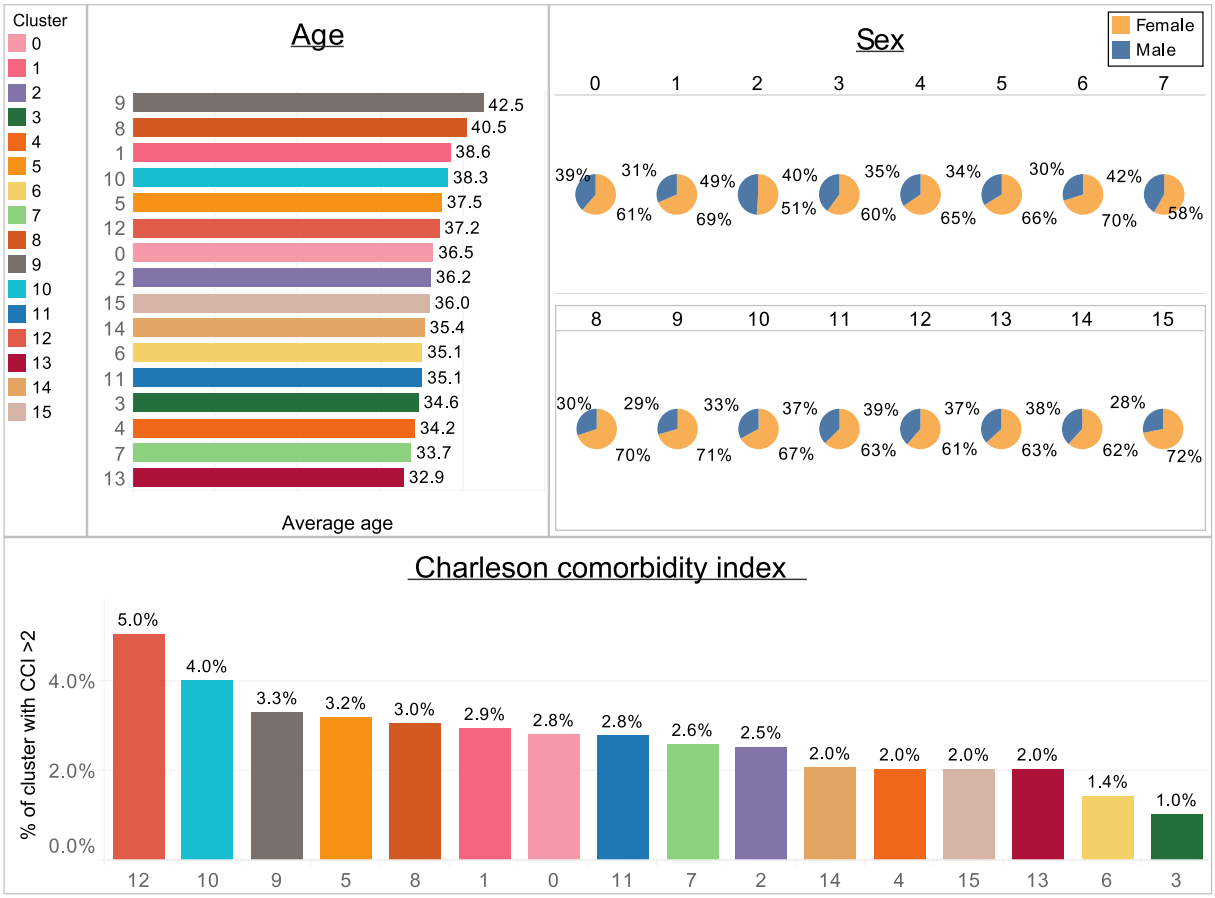

Fig. 7 Comparison of age, sex, and CCI across clusters. CCI, Charlson comorbidity index

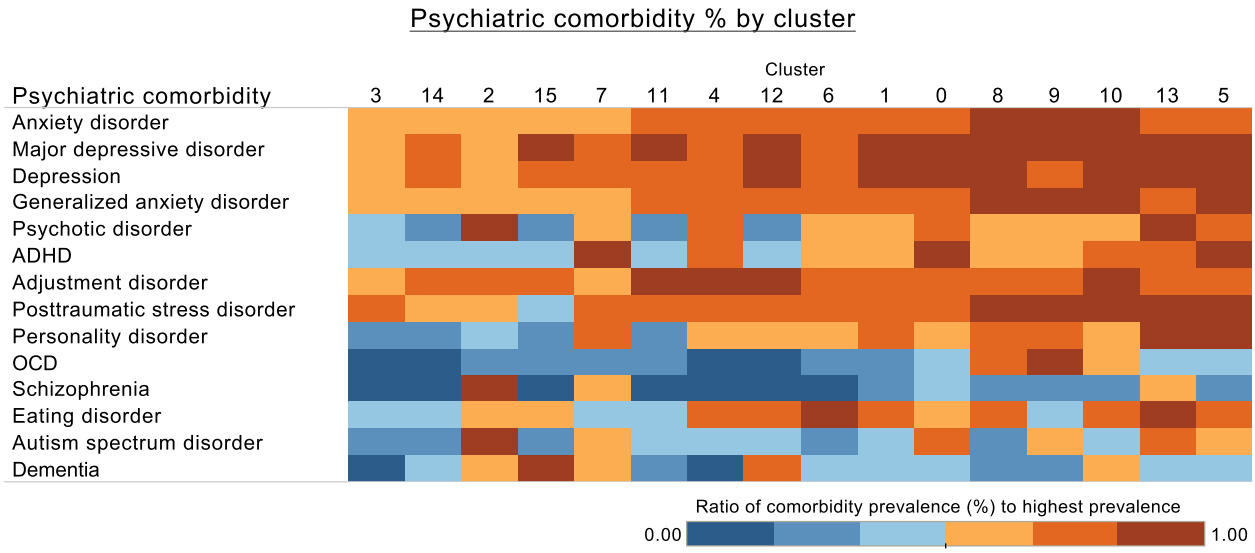

Fig. 8 Ratio of comorbidity prevalence at baseline. The heatmap shows each cluster's psychiatric comorbidity prevalence (percentage of patients with each comorbidity), normalized to the highest prevalence for each comorbidity across clusters. ADHD, attention-deficit/hyperactivity disorder; OCD, obsessive–compulsive disorder

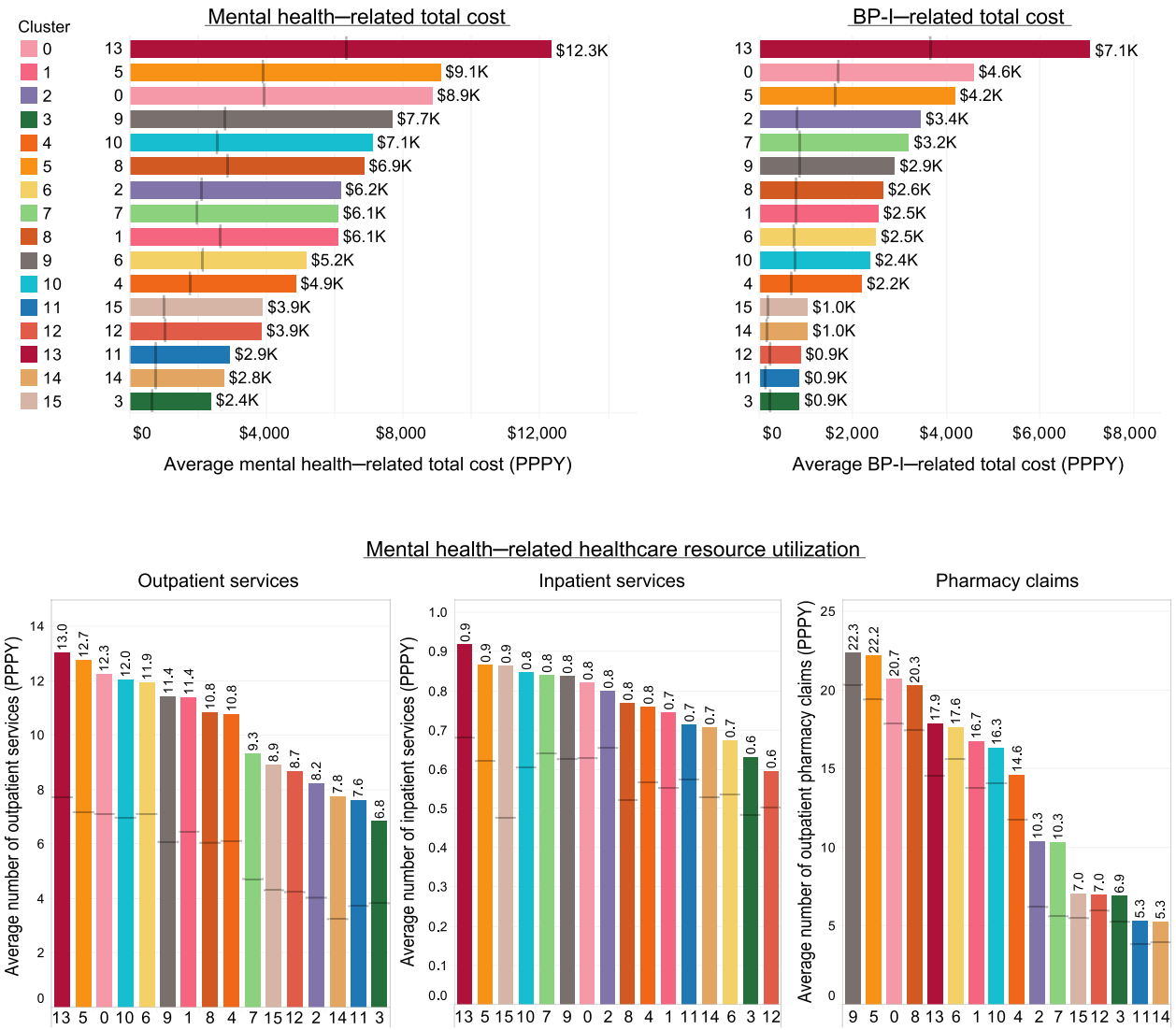

Fig. 9 Average total mental health– and BP-I–related costs and mental health–related resource use by cluster. Total cost is the sum of inpatient, outpatient, and pharmacy costs. Gray lines indicate the median value for each parameter per cluster. BP-I, bipolar I disorder; PPPY, per patient per year

sequence length, as well as the ability to set a minimum cluster size. These characteristics allowed clustering based on clinically relevant events from a diverse US commercially insured population with a complex psychiatric condition, BP-I. With adjustments to preprocessing steps, the approach could be adapted to other BP-I event types or other disease indications. Further assessment and external validation should be conducted using larger sample sizes and different databases. Future research should include performance comparison against other established clustering algorithms, such as the two-step hidden Markov model approach [11], the Temporal Needleman-Wunsch approach [12], and an algorithm for medical treatment process modeling [29]. However, many algorithms may not be compatible with the diversity of psychiatric event types or large-scale datasets.

DOCKKES has the potential to enable large-scale treatment journey analyses by unraveling complexity to create coherent patient clusters. As a natural next step, DOCKKES-identified clusters can be used for process mining, which can systematically compare real-world treatment journeys to clinical treatment guidelines. Outcomes, such as HRU and cost, could be analyzed to identify suboptimal versus optimal treatment approaches. Other useful analytics could include the prediction of future events or the identification of intervention points that can be

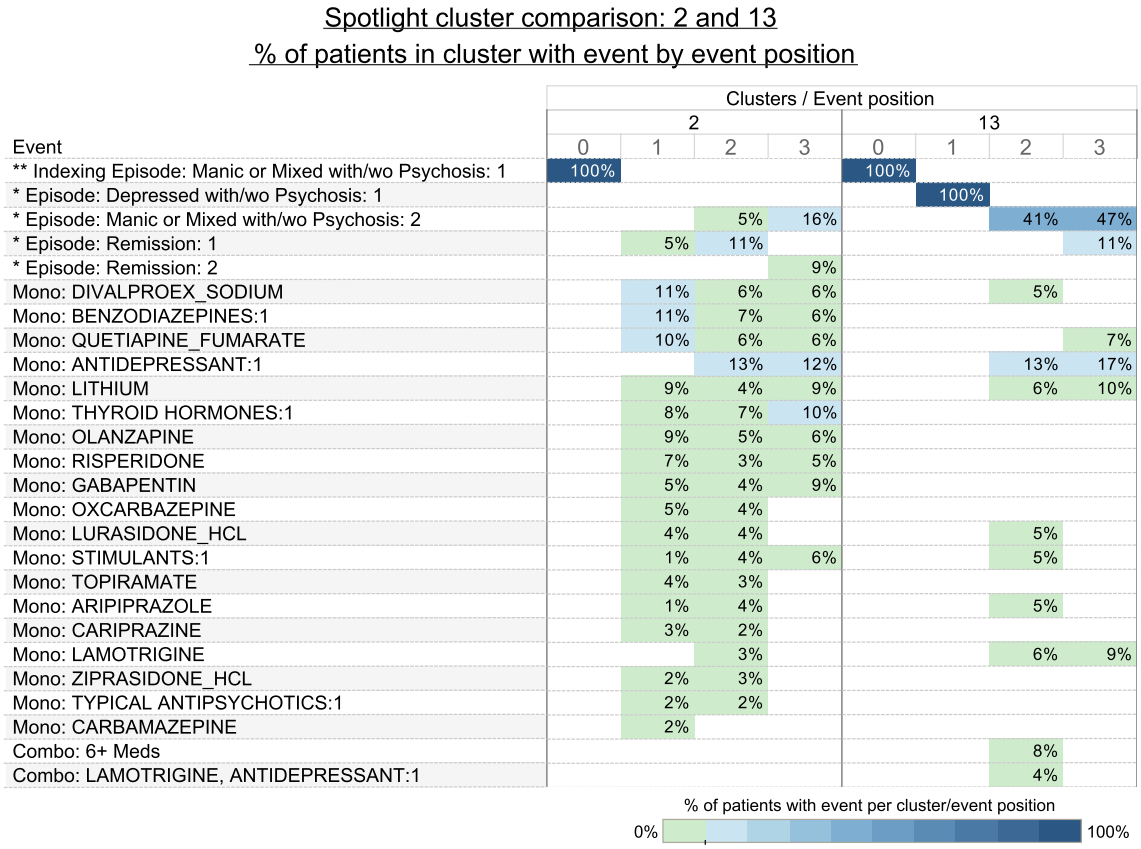

Fig. 10 Percentage of patients with event per cluster and event position in clusters 2 and 13. The first four event positions are shown in columns 0, 1, 2, and 3

used to improve outcomes. It would also be interesting to develop a temporal approach (eg, as in the study by Jin et al. 2008 [7]) compatible with psychiatry-relevant events to investigate the importance of event timing in addition to sequence.

4.1 Limitations

This study had several limitations. It is possible that treatment journeys are differentiated by uncaptured factors, such as geographic region (ie, statewide practice approaches to BP-I). Claims data may not capture all events that are meaningful to a patient's journey, such as those related to socioeconomic determinants of health. For example, a patient may be prescribed a treatment yet not fill the prescription, perhaps due to out-of-pocket costs. Nonadherence to medication dosing is also not captured. The analysis design and use of claims data may have introduced selection bias; particularly in the context of our BP-I sample, patients remained insuredthroughout our study, which may not be generalizable to all real-world psychiatry populations. However, the Marketscan Database is the largest real-world US claims database; therefore, the study's large sample size supports meaningful clusters in the context of BP-I treatment patterns and potential clinical hypotheses regarding patient profiles based on the resulting clusters. In this exploratory analysis, simple descriptive comparisons were made between cohorts. Further analyses are needed to assess statistical differences in key outcomes (HRU and cost) across clusters and adjust for variables like demographics, clinical characteristics, and length of follow-up.

5 Conclusion

In this proof-of-concept study, the novel DOCKKES clustering approach supported an exploratory analysis of BP-I treatment journeys by flexibly integrating information from diverse sequences of episode and treatment events. DOCKKES produced differentiated, potentially clinically relevant treatment journey clusters. These clusters suggested guideline-incongruent patterns and ultimately may be useful in discovering associations between treatment processes and outcomes such as HRU and cost. Additionally, DOCKKES has the potential to provide insight into other mental health conditions with complex diagnostic and treatment patterns.

Abbreviations

ADHD	Attention-deficit/hyperactivity disorder
AWP	Average wholesale price
BP-I	Bipolar I disorder
BPE	Bipolar episode
CCI	Charlson comorbidity index
CDHP	Consumer-directed health plan
CE	Continuous enrollment
COMP	Comprehensive
DOCKKES	Divisive Optimized Clustering using Kernel K-means for Event Sequences
EPO	Exclusive provider organization
HCl	Hydrochloride
HMO	Health maintenance organization
HRU	Healthcare resource utilization
ICD	International Classification of Diseases
LOT	Line of therapy
OCD	Obsessive–compulsive disorder
POS	Point of service plan
POS-C	Point of service plan with capitation
PPO	Preferred provider organization
PPPY	Per patient per year
SD	Standard deviation
SNRI	Serotonin and norepinephrine reuptake inhibitor
SSRI	Selective serotonin reuptake inhibitor
TCA	Tricyclic antidepressants
TNW	Temporal Needleman-Wunsch

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40708-025-00258-x.

Supplementary material 1.

Acknowledgements
The authors would like to thank Dr. Amanda Harrington for her support with this project. Medical writing and editorial support were provided by Ellen Labuz, PhD, from Prescott Medical Communications Group, a Citrus Health Group, Inc., company (Chicago, Illinois), as well as Stuart Rutlen, Valerie Moss, and Rosie Henderson from Onyx (London, United Kingdom) and funded by AbbVie.

Author contributions
All authors were involved with the study design, analysis, and interpretation of data. All authors participated in the writing, editing, and critical revision of intellectual content, as well as approval of the final version of this manuscript. All authors met the International Committee of Medical Journal Editors authorship criteria and agreed to be accountable for all aspects of the work.

Funding
AbbVie funded this analysis and participated in the design, research, analysis, data collection, interpretation of data, reviewing, and approval of the publication. All authors had access to relevant data and participated in the drafting, review, and approval of this publication. No honoraria or payments were made for authorship.

Data availability
DOCKKES code, written in platform-independent Python 3, is available from https://github.com/abbvie-external/DOCKKES. Data from the MarketScan Commercial Database are proprietary to Merative and, thus, are not publicly available without a licensing agreement.

Declarations

Competing interests
All authors are employees of AbbVie and may hold stock.

Received: 23 August 2024 Accepted: 27 April 2025
Published online: 22 May 2025

References

1. Goldberg JF (2008) Optimizing treatment outcomes in bipolar disorder under ordinary conditions. J Clin Psychiatry 69(Suppl 3):11–19
2. Gorwood P (2006) Meeting everyday challenges: antipsychotic therapy in the real world. Eur Neuropsychopharmacol 16(Suppl 3):S156–S162. https://doi.org/10.1016/j.euroneuro.2006.06.002
3. Jain R, Kong AM, Gillard P, Harrington A (2022) Treatment patterns among patients with bipolar disorder in the United States: a retrospective claims database analysis. Adv Ther 39:2578–2595. https://doi.org/10.1007/s12325-022-02112-6
4. Sen J, Tonkin A, Varigos J et al (2021) Risk stratification of cardiovascular complications using CHA(2)DS(2)-VASc and CHADS(2) scores in chronic atherosclerotic cardiovascular disease. Int J Cardiol 337:9–15. https://doi.org/10.1016/j.ijcard.2021.04.067
5. Jia X, Baig MM, Mirza F, GholamHosseini H (2019) A cox-based risk prediction model for early detection of cardiovascular disease: identification of key risk factors for the development of a 10-year CVD risk prediction. Adv Prev Med 2019:8392348. https://doi.org/10.1155/2019/8392348
6. Aspland E, Gartner D, Harper P (2019) Clinical pathway modelling: a literature review. Heal Syst 10:1–23. https://doi.org/10.1080/20476965.2019.1652547
7. Jin HW, Chen J, He H et al (2008) Mining unexpected temporal associations: applications in detecting adverse drug reactions. IEEE Trans Inf Technol Biomed 12:488–500. https://doi.org/10.1109/titb.2007.900808
8. Sommers D, Menkovski V, Fahland D (2021) Process discovery using graph neural networks. pp 40–47.
9. Rovani M, Maggi FM, de Leoni M, van der Aalst WMP (2015) Declarative process mining in healthcare. Expert Syst Appl 42:9236–9251. https://doi.org/10.1016/j.eswa.2015.07.040
10. Song M, Günther CW, van der Aalst WMP (2009) Trace clustering in process mining. Springer, Berlin Heidelberg, pp 109–120
11. Najjar A, Reinharz D, Girouard C, Gagné C (2018) A two-step approach for mining patient treatment pathways in administrative healthcare databases. Artif Intell Med 87:34–48. https://doi.org/10.1016/j.artmed.2018.03.004
12. Rama K, Canhão H, Carvalho AM, Vinga S (2019) AliClu—temporal sequence alignment for clustering longitudinal clinical data. BMC Méd Inform Decis Mak 19:289. https://doi.org/10.1186/s12911-019-1013-7
13. Wu W, Yan J, Yang X, Zha H (2022) Discovering temporal patterns for event sequence clustering via policy mixture model. IEEE Trans Knowl Data Eng 34:573–586. https://doi.org/10.1109/tkde.2020.2986206
14. Bampa M, Miliou I, Jovanovic B, Papapetrou P (2024) M-ClustEHR: a multimodal clustering approach for electronic health records. Artif Intell Med 154:102905. https://doi.org/10.1016/j.artmed.2024.102905
15. Wu C-P, Sleiman J, Fakhry B et al (2024) Novel machine learning identifies 5 asthma phenotypes using cluster analysis of real-world data. J Allergy Clin Immunol Pr 12:2084-2091.e4. https://doi.org/10.1016/j.jaip.2024.04.035
16. Chin S, Collins JE (2024) Clustering methods in rheumatic and musculoskeletal diseases research: an educational guide to best research practices. J Rheumatol 51: jrheum. 2024–0519. https://doi.org/10.3899/jrheum.2024-0519
17. Bernini S, Valcarenghi A, Ballante E et al (2025) A data-driven cluster analysis to explore cognitive reserve and modifiable risk factors in early phases of cognitive decline. Sci Rep 15:4616. https://doi.org/10.1038/s41598-025-88340-6
18. Gao CX, Dwyer D, Zhu Y et al (2023) An overview of clustering methods with guidelines for application in mental health research. Psychiatry Res 327:115265. https://doi.org/10.1016/j.psychres.2023.115265
19. Shutaywi M, Kachouie NN (2021) Silhouette analysis for performance evaluation in machine learning with applications to clustering. Entropy 23:759. https://doi.org/10.3390/e23060759
20. Debolina P, Chakraborty S, Das S, Xu J (2022) Implicit annealing in kernel spaces: a strongly consistent clustering approach. IEEE Trans Pattern Anal Mach Intell Pp: 5862–5871. https://doi.org/10.1109/tpami.2022.3217137

21. McIntyre RS, Alda M, Baldessarini RJ et al (2022) The clinical characterization of the adult patient with bipolar disorder aimed at personalization of management. World Psychiatry 21:364–387. https://doi.org/10.1002/wps.20997
22. Yatham LN, Kennedy SH, Parikh SV et al (2018) Canadian network for mood and anxiety treatments (CANMAT) and international society for bipolar disorders (ISBD) 2018 guidelines for the management of patients with bipolar disorder. Bipolar Disord 20:97–170. https://doi.org/10.1111/bdi.12609
23. Merative (2023) MarketScan research databases for life sciences researchers. https://www.merative.com/content/dam/merative/documents/brief/marketscan-research-databases-for-life-sciences-researchers.pdf
24. Vogt V, Scholz SM, Sundmacher L (2018) Applying sequence clustering techniques to explore practice-based ambulatory care pathways in insurance claims data. Eur J Public Heal 28:214–219. https://doi.org/10.1093/eurpub/ckx169
25. Halkidi M, Batistakis Y, Vazirgiannis M (2001) On clustering validation techniques. J Intell Inf Syst 17:107–145. https://doi.org/10.1023/a:1012801612483
26. Tzortzis GF, Likas AC (2009) The global kernel k-means algorithm for clustering in feature space. IEEE Trans Neural Netw 20:1181–1194. https://doi.org/10.1109/tnn.2009.2019722
27. Cartis C, Roberts L, Sheridan-Methven O (2022) Escaping local minima with local derivative-free methods: a numerical investigation. Optimization 71:2343–2373. https://doi.org/10.1080/02331934.2021.1883015
28. Florida University of South (2020) 2019–2020 Florida best practice psychotherapeutic medication guidelines for adults. https://floridabhcenter.org/wp-content/uploads/2021/04/2019-Psychotherapeutic-Medication-Guidelines-for-Adults-with-References_06-04-20.pdf
29. Yang L, Kang G, Zhang L (2021) Medical treatment process modelling based on process mining and treatment patterns. China Commun 18:332–349

Publisher's Note