Medal: a patient similarity metric using medication prescribing patterns.

Arturo Lopez Pineda, PhD[1], Armin Pourshafeie, MS[1,2], Alexander Ioannidis, PhD[1], Collin McCloskey Leibold, MS[3], Avis Chan, MD[3], Jennifer Frankovich, MD, MS[3], Carlos D. Bustamante, PhD[1,4,5,*], Genevieve L. Wojcik, PhD[1,6,*]

1. Department of Biomedical Data Science, Stanford University, CA, USA
2. Department of Physics, Stanford University, CA, USA
3. Department of Pediatrics, Division of Allergy, Immunology and Rheumatology, Stanford University, CA, USA
4. Department of Genetics, Stanford University, CA, USA
5. Chan Zuckerberg Biohub, San Francisco, CA, USA
6. Department of Epidemiology. Bloomberg School of Public Health. Johns Hopkins University, Baltimore, MD, USA

*Corresponding authors

Address correspondence to:

Genevieve L. Wojcik, PhD (gwojcik1@jhu.edu)
Department of Epidemiology. Bloomberg School of Public Health.
Johns Hopkins University
Baltimore, Maryland, United States

Carlos D. Bustamante, PhD (cdbustam@stanford.edu)
Department of Biomedical Data Science. Stanford University School of Medicine
Stanford, California, United States

Abstract

Patient similarity offers an exciting potential to improve our understanding of treatment patterns. Medication history is a valuable source of information given the clinical considerations taken during the physician's prescription process. However, available similarity methods do not consider timeliness of event occurrence in the longitudinal care of patients.

In this observational cohort study, we propose an event alignment algorithm, *Medal*, which uses a dynamic programming approach for pairwise alignment of medication histories. To test our algorithm, we obtained the medication usage from a cohort of patients with pediatric acute-onset neuropsychiatric syndrome (PANS). After calculating a distance metric with *Medal*, we compute a hierarchical clustering and explore the most appropriate number of clusters.

We identified four clusters in PANS with distinct medication usage histories, driven primarily by penicillin. We foresee that our algorithm could be used to identify clusters in other syndromes treated with multiple medication regimens.

Keywords: Cluster Analysis; Medical Informatics; Patient similarity; Longitudinal Studies; Polypharmacy

1. Background

In precision medicine, patient similarity is an emerging concept that aims to estimate a numerical distance between components of patient data[1], in order to find groups of patients (clusters) that share short distances in those components; and ultimately, use those clusters in predictive modeling tasks[2]. Patient similarity measurements (PSM) typically[3,4] employ distance functions that consider geometrical space (e.g. Euclidean, Manhattan, Mahalanobis, etc.). Novel approaches have been developed to estimate patient similarity which are not geometric-based, for example using machine learning models to estimate the distance between patients with decision trees[5] or random forests[6]; or the use of ontologies to extract hierarchically related diagnosis[7].

The vector of features used for PSM include patient information like genetics[8], demographics and population characteristics[9], prescriptions and lab tests[4,10], medical billing codes[11], and even clinical narratives that get processed with natural language tools to extract features from the text[12]. However, the temporal dimension has not been sufficiently explored. Medication usage, for example, is a longitudinal source of patient information which is generated and continuously recorded in the electronic medical record (EMR) over the course of care delivery.

In cohorts of patients with polypharmacy treatment, the use of multiple medications, it is important to understand if there are groups of patients that share similar patterns of usage and what are the differences between these groups of patients. Polypharmacy is common in older patients with multimorbidity[13]; and it is associated with adverse outcomes including mortality and adverse drug reactions[14], increased length of stay in hospital and readmission to hospital soon after discharge[15]. However, polypharmacy can also occur in children and adolescent patients with psychiatric diseases[16], and other non-elder adults with complex chronic syndromes such as lupus[17], human immunodeficiency virus infection[18], ischemic and respiratory diseases[19], or cancer[20].

These diseases are often multifactorial, with physicians treating related, but separate, symptoms and pathologies. One of these syndromes is the pediatric acute-onset neuropsychiatric syndrome (PANS).

The pediatric acute-onset neuropsychiatric syndrome (PANS)

The PANS clinical presentation[21] is characterized by abrupt-onset of obsessive-compulsive disorder (OCD) and/or food restriction, along with at least two other severe neuropsychiatric symptoms from the following categories: anxiety; mood lability or depression; irritability, oppositionality, or rage; behavioral regression; deterioration in school performance/cognitive difficulties; sensory or motor abnormalities; and somatic symptoms like sleep disturbances or enuresis. Patients typically experience a relapsing-remitting course in which disease flares are interspersed with remissions[22]. In some cases, the disease course is chronic, when the patient's neuropsychiatric status does not return to baseline.

Some evidence suggests that PANS has an inflammatory or autoimmune etiology that may be triggered by an infection[23,24]; therefore, multidisciplinary clinics are well-positioned to care for patients with PANS[25]. Although treatment data are critically lacking, interim guidelines suggest using antibiotics to treat or prevent infections, immunomodulatory therapies to manage inflammation, and psychiatric medications supplemented with cognitive behavioral therapy to treat PANS[26]. The heterogeneity and complexity of PANS presentation, clinical course, treatment, insurance status, and irregular follow-up make it difficult to compare treatment course across patients and patient-groups, necessitating a novel method to cluster patients while taking into account temporality.

Clinical significance

In this study we developed an algorithm for clustering patients based on their medication usage. Our algorithm creates a patient similarity measurement considering the longitudinal nature of medication usage. The algorithm, called *Medal*, follows a dynamic programming approach to perform a pairwise alignment of the medication usage. The necessary steps to align two pairs of medication histories is then used to build clusters of similar patients. With the use of this algorithm and the medication usage histories, we aim to discover groups of patients sharing similar journeys in their patient care.

2. Results

Clinical characteristics

Patient characteristics in this study were evenly distributed by sex, and heavily skewed by self-reported ethnicity (mostly non-Hispanic white) patients. The age at first neuropsychiatric symptoms occurred on average between 7-8 years old, with a rapid intake by the PANS clinic (patients seen > 4 months after psychiatric symptom onset were excluded). Detailed description of the clinical characteristics of this cohort can be seen in Table 1.

Table 1. Clinical characteristics of 43 consecutive pre-pubescent patients with new-onset PANS.

Characteristic	PANS
Overall	**N = 43**
Sex	
Female	23 (53.49%)
Male	20 (46.51%)

Age at first neuropsychiatric symptom onset (years)	mean = 7.77 SD = 2.35
Age at first clinic visit (years)	mean = 7.93 SD = 2.39
Ethnicity*	
Non-Hispanic White	39 (90.7%)
Non-Hispanic Asian	4 (9.3%)
Hispanic / Latino	4 (9.3%)
Other / unknown	4 (4.65%)
PANS symptoms at first presentation to the clinic	
Obsessive Compulsive Disorder (OCD)	36 (83.72%)
Food intake problems	18 (41.86%)
Anxiety / phobia	37 (86.05%)
Emotional lability / depression / suicidal ideation	33 (76.74%)
Aggression / irritability / opposition	33 (76.74%)
Cognitive problems	9 (20.93%)
Inattention / deterioration in school	24 (55.81%)
Behavioral regression	27 (62.79%)
Sensory amplification	19 (44.19%)
Sleep disturbances	26 (60.47%)
Urinary problems	14 (32.56%)
Motion / vocal tics	24 (55.81%)

SD = standard deviation. *Ethnicity is a self-reported item on the patient questionnaire and more than one category could be selected.

The medication list included seven categories of drugs: 1) penicillins, 2) cephalosporins, 3) macrolides, 4) non-steroidal anti-inflammatory drugs (NSAIDs), 5) corticosteroids (oral and intravenous), 6) antibodies, and 7) disease-modifying anti-rheumatic drugs (DMARDs). For each patient, a detailed medication history was collected and included initiation date and intake duration. This input was then used by Medal to create an edit distance matrix, from which a hierarchical cluster was built. At the time of this analysis, psychiatric medications taken were not yet input in the PANS Redcap research database.

Number of Clusters

To select the optimal number of clusters, we applied three evaluations using the elbow, silhouette[27], and gap statistic[28] methods, shown in Figure 1. On a graphical inspection of the elbow method, there are elbows in the plot at k={2, 4, 6, 8, 9, 10}, however, the most evident elbow was deemed to be k=4. For the Silhouette method, the maximum value was at k=9 (the highest rated), however, on a greedy approach, the highest value would

be k=2, followed by k=4. Finally, for the gap statistic method, after running 500 bootstrap, the best value was k=2. Given these methods we selected k=2 and k=4.

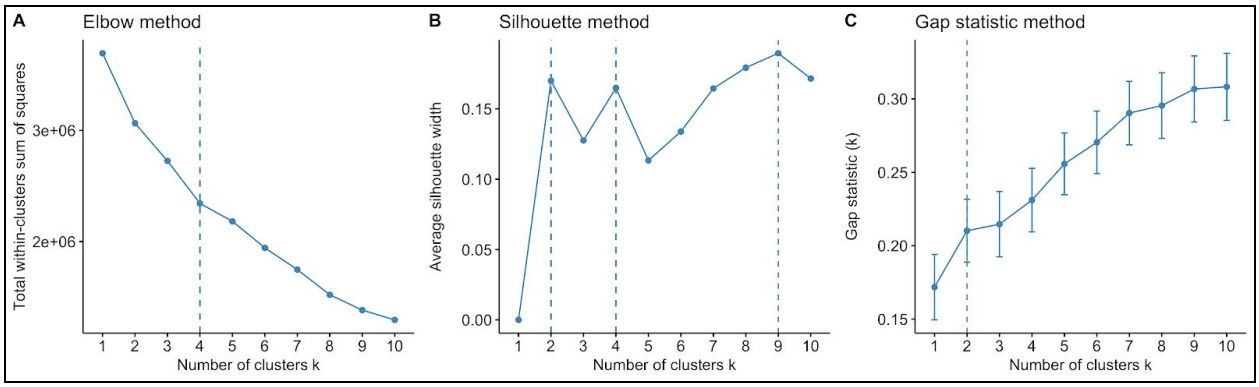

Figure 1. Analysis of optimal number of clusters. A) Elbow method, B) Silhouette method, C) Gap statistic method.

Cluster assignment

The hierarchical clustering, shown in Figure 2, provided more information about the best number of clusters to investigate. For the remaining of the analysis we will be using k=4, with the caveat that they are really two clusters formed by two subclusters. The multidimensional scaling (MDS) plots shown in panels B and C of Figure 2 show that the first component can differentiate between the two super clusters (first super cluster is formed by subclusters 1-yellow and 3-green, while the second super cluster is formed by subclusters 2-red, and 4-blue). In MDS space, the second and third components provides small differences in-between subclusters. Additional information is provided in the Supplementary Material regarding cluster assignment evaluation.

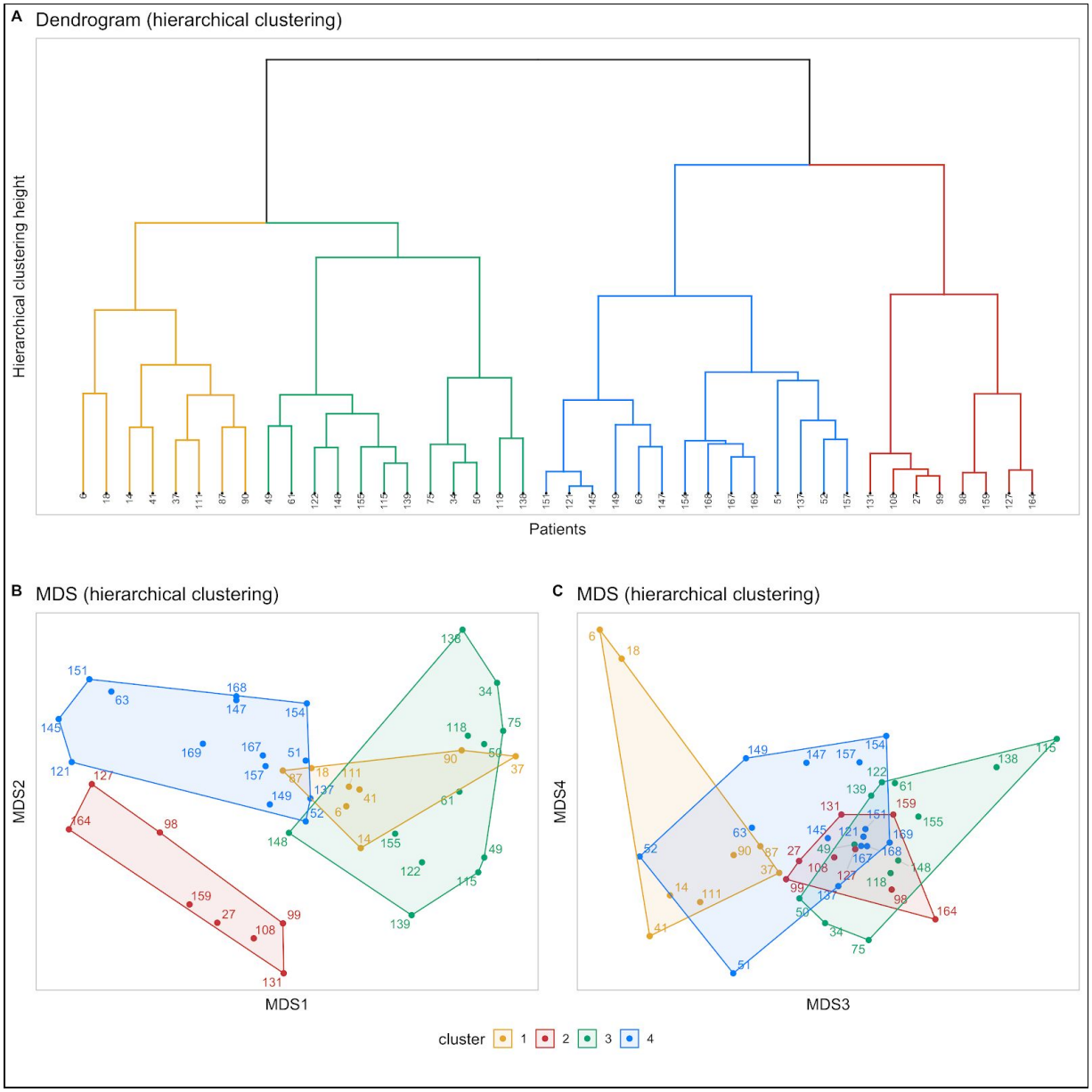

Figure 2. Cluster assignment for hierarchical clustering using k = 4. A) Dendrogram, B) MDS components 1 and 2, C) MDS components 2 and 3.

Cluster characteristics

We also assessed the qualitative value of the clustering assignment, based on the patient similarity metric calculated using *Medal*. The medication usage by cluster and medication class is shown in Figure 3. The figure is censored at two years of follow-up,

and the index date used to align all medication histories is the first visit to the PANS clinic. Each medication class varies from none of the patients taking that drug to all of the patients taking the drug, which was grouped by month to facilitate the visualization. Subclusters 1 and 3 (same supercluster) are characterized by moderate to heavy usage across most of the drug classes, being differentiated by lack of macrolide usage in subcluster 3. Subclusters 2 and 4 (same supercluster) are characterized with moderate to low usage of medications, with the exception of penicillin in subcluster 2 and NSAIDs in subcluster 4.

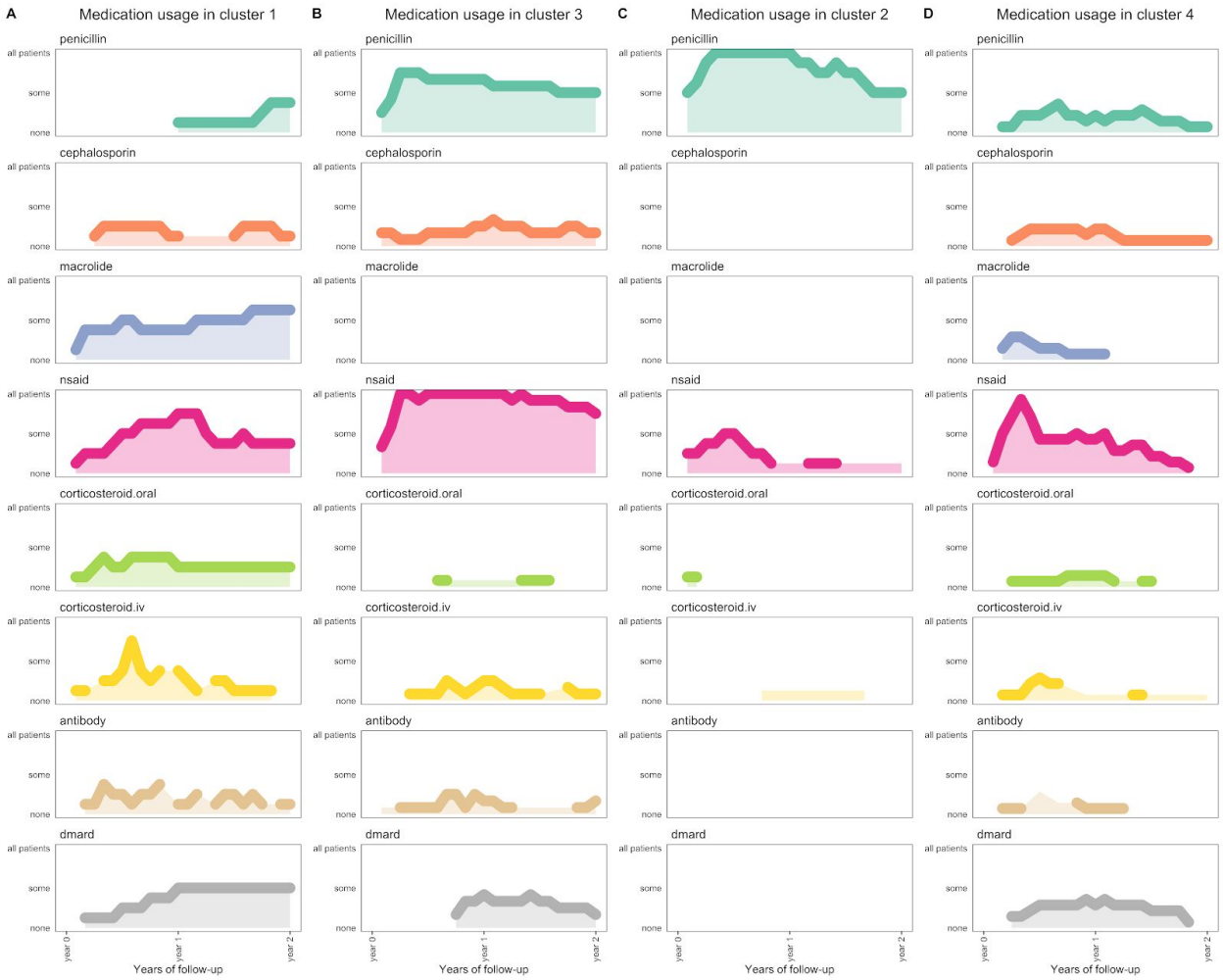

Figure 3. Drug usage of patients in four clusters

In the PANS cohort, like in many other psychiatric syndromes, evaluating outcomes is a complex task which requires the use of scales and assessment scores that are subjective in nature. The Stanford PANS clinic utilizes two parent rated impairment scales that have been validated in this patient population: the Global Impairment Scale[29] and the Caregiver Burden Inventory[30]. Figure 4 shows the evolution in time of these scales by cluster assignment, for other scales please refer to the Supplementary Material.

The outcome data (Global Impairment and Caregiver Burden) are skewed by the fact that the frequency of follow-up clinic visits and corresponding completed questionnaires trends with the severity of the psychiatric symptoms (i.e. when patients are highly symptomatic, they come to the clinic weekly for psychiatric medication titration and therapy, but when the child improves/resolves their psychiatric symptoms, it is difficult to get the family to return to clinic.

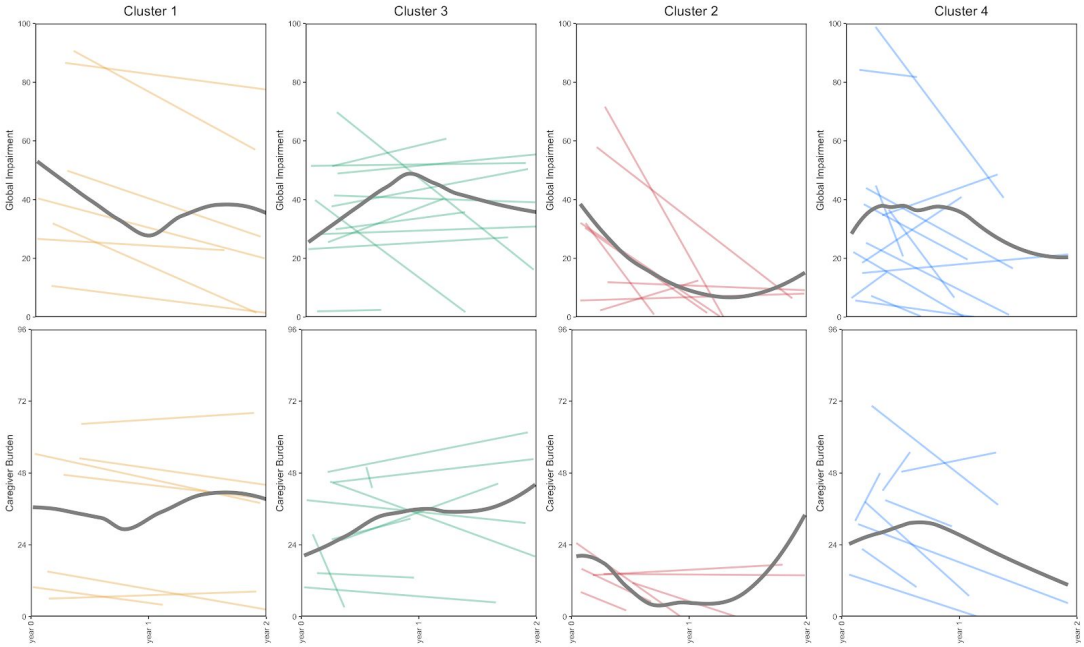

Figure 4. Overall impairment scales by cluster. First row, Global Impairment Scale. Second row, Caregiver Burden Inventory. Fine lines show the linear trend for each patient in that cluster, while the bolded line shows the polynomial regression fitting for all patients in that cluster.

Upon performing an analysis of variance (ANOVA), we found in Table 2 that there is significant heterogeneity in trends between individuals for both Global Impairment (P_{ANOVA}=0.013) and Caregiver Burden (P_{ANOVA}=0.035). We note that clusters 1, 2, and 4 demonstrate on average decreased Caregiver Burden Inventory and decreased Global Impairment Scale over time, with the exception of cluster 3. However, the sample sizes are too small to draw conclusions from these trends.

Table 2. Slopes for Caregiver Burden Inventory and Global Impairment scores over time by cluster.

	Global Impairment Score				**Caregiver Burden**			
	Intercept	*Slope*	*Std. Error*	*P-value*	*Intercept*	*Slope*	*Std. Error*	*P-value*
Cluster 1	50.00	-10.85	2.75	<0.001	37.30	-4.51	1.88	0.019
Cluster 2	31.45	-16.65	6.52	0.036	16.13	-5.71	4.91	0.321
Cluster 3	36.53	-0.168	3.68	0.965	29.29	0.33	3.54	0.927
Cluster 4	36.80	-12.85	6.30	0.063	39.37	-16.01	4.27	0.100
				ANOVA P-value: **0.013**				ANOVA P-value: **0.035**

3. Discussion

Here we present a novel method for the alignment of patient history to identify distinct clusters of medication usage, *Medal*, and the application to a pediatric cohort.

On algorithm efficiency.

Our approach relies on constructing a symmetric matrix of the pairwise distances between patients. The dynamic programming approach used for constructing the distance between each pair scales as the square of the time period under consideration. Therefore, construction of the distance matrix for N patients taking D drugs over a period of length T takes $O(DN^2T^2)$ time and uses $O(N^2+T^2)$ memory. In cases where the exact intervals and number of days that the drug is administered is not crucial the

time periods can be collapsed (e.g. from days to weeks) in order to increase computational and memory efficiency. Our implementation can be accessed through the repository: https://github.com/bustamante-lab/medal

Limitations

Our study was limited by the availability of data in the PANS Redcap research database. The small number of patients selected for this study may have also biased our findings. The selection of patients was reduced from an initial cohort of 305 patients seen in the Stanford PANS clinic to 43 patients who met the strict PANS criteria and who were "new-onset" at the time of clinic entry. Furthermore, we limited the timespan of medication history to only the first two years of treatment, to make the patient comparisons possible. The strict study criteria and limited time frame increased comparability between patients to find patterns related to medication usage immediately after diagnosis and treatment initiation. Additionally, we grouped medication categories instead of treating each drug independently to increase power.

Clinical importance of cluster assignment.

Our strategy can help identify clusters to characterize the patient population, however, it is important to make the distinction that assignment to a cluster should not be considered an association without further investigation. This approach is merely a way to generate new hypotheses that could be further investigated by the clinical team. In a polypharmacy context, this approach could further be used to better understand treatment patterns in the clinic.

Based on medication history, including initiation and cessation dates, we identify four apparent clusters in the PANS cohort. It is interesting to note that in the representation of Figure 3, the superclusters 1-3 and 2-4 seem to differ in the amount of medication usage, with the former supercluster being under more medication than the latter. Less

perceptible differences can be visually observed in Figure 3, when trying to find distinctions in the subclusters.

Subcluster 1 had a peak of intravenous corticosteroid, followed by a steady increase in use of several other medications. This pattern is consistent with a more severe phenotype since the clinicians in the PANS clinic reserve IV corticosteroids for the most severe cases. This is supported by the outcomes data (Fig 4) which indicate that Subcluster 1 has the highest Global Impairment and Caregiver Burden at clinic entry and throughout the 2 years of study except perhaps after the peak use of IV corticosteroids.

Subcluster 3 had the lowest Global Impairment scores (despite high Caregiver Burden) and had a constant heavy use of NSAIDs which may represent a subgroup of patients who respond to NSAIDs but rely on the constant use of NSAIDS to suppress symptoms. NSAIDs are continued long term only in the subset of patients with PANS who have recrudescence of symptoms when the NSAID dose is lowered and in patients who also develop arthritis. However, this group seemed to worsen over time (especially in the first year) which appears to be associated with the eventual addition of more aggressive immunomodulation. The low Global Impairment at clinic entry may have led to the decision not to use corticosteroids in the initial treatment.

Subclusters 2 and 3 are characterized by heavy penicillin use which may identify this group as the subsets where there was evidence for a streptococcal infection coincident with the onset of the psychiatric illness (an association seen in epidemiological studies[21]) and likely reflects the clinicians attempt to use penicillin as prophylaxis against streptococcus.

Subcluster 2 has a high Global impairment and low Caregiver Burden at onset which rapidly improves over the first 1-1.5 years. This patient group has the highest penicillin usage of the four groups. Late in year two, Global Impairment and Caregiver burden

increase which likely reflects relapse and appears to be coincident with decreasing penicillin usage.

Conclusion and Future work

The field of patient similarity is expanding with the inclusion of novel sources of data in electronic format. In this study, we have shown that *Medal* is capable of providing a reliable similarity metric that can lead to the investigation of new hypotheses for a complex syndrome like PANS. The medication histories of other cohorts with a high pharmacological burden will play an important role in our understanding of their treatment patterns. We envision Medal, as the first of a long list of novel patient similarity algorithms that could incorporate the longitudinal nature of medication usage.

4. Methods

This is a retrospective longitudinal cohort study, using patient/parent questionnaires and ledgers collected routinely as part of clinical care. The objective was to understand patterns of medication usage in patients. The STROBE[31] checklist is provided in Supplementary Material.

Clinical setting and population

The study took place at the Stanford PANS clinic located in California, which is a multidisciplinary clinic staffed by practitioners of various disciplines (psychiatry, psychology, primary care, rheumatology, immunology, otolaryngology, and neurology), as well as a social work psychotherapist and an education specialist.

The data used in this study were generated during clinical care, but we retrospectively collected information on those patients seen between the opening of the clinic in November 2012, and the patient identification date for this study, January 2018. In this period there were 305 patients, out of which we excluded: those whose parents refused

consent for the study (N = 4), those patients who did not meet the strict PANS criteria[21] (N = 97), had fewer than three visits to the PANS clinic (N = 36), older than 12 years at clinic entry (N = 60) and presented to the clinic more than 4 months after onset of psychiatric symptoms (N = 65). The final cohort includes 43 pre-pubescent new-onset patients, as shown in Figure 5. One patient was dropped from the primary analysis because the only medication was taken after the first two years at clinic.

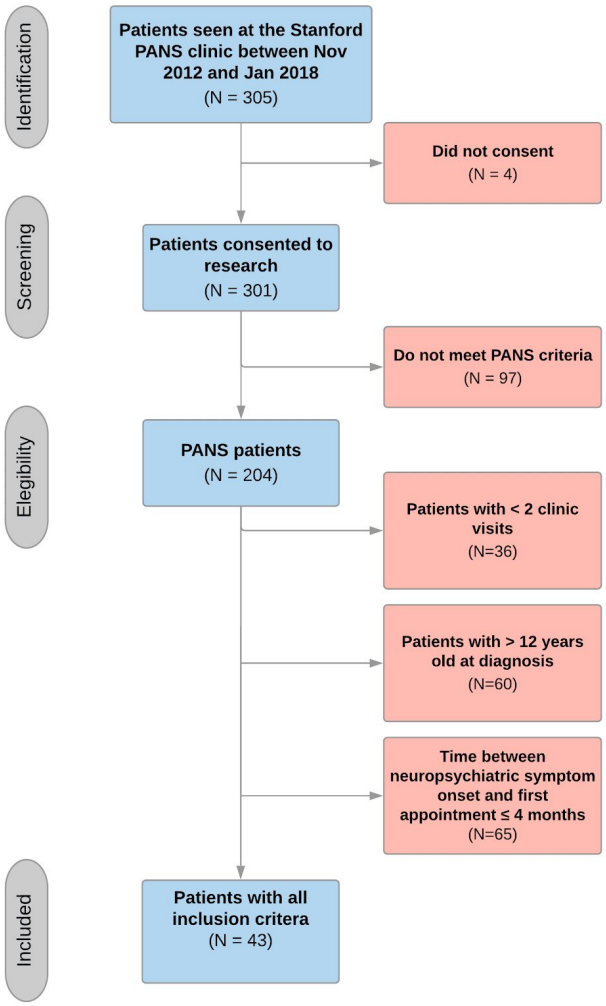

Figure 5. PANS inclusion and exclusion criteria.

We only included patients with established care with the clinic (at least three visits) because the first few visits aim to understand the history and disease course, examine patients and counsel parents about different treatment options. We used a cut-off age

of 12 years to include only pre-pubertal children, as hormones might play a role in psychiatric symptoms and behaviors[32]. We restricted patients to those with such a short time difference between PANS onset and first PANS clinic appointment because we wanted to analyze patients with new onset PANS, for whom we had more complete information close to the beginning of their PANS illness.

Data sources

We collected data on medical treatment for this neuropsychiatric syndrome, including antibiotics and immunomodulators from the electronic medical record (EMR) using a keyword search method[33], and made a final update of medication history in mid-October 2018. We excluded short courses (less than 21 days) of antibiotics for acute infections, NSAID taken if needed by patients, and psychiatric medications as the aim of this analysis is to study the similarity of using medical (non-psychiatric/non-psychological) therapies in this group of patients. Our keyword search method is outlined in Table 3.

Table 3. Medication list and keywords (underlined) used in the EMR search-box.

Class	Generic name	Brand name	Common use	Freq	Admin.
Antibiotic (Penicillin)	Penicillin G	various	Streptococcal infection and prophylaxis against strep	Once every 3-4 weeks	IM
	Penicillin V	various	*Idem*	BID to QID	Oral
	Amoxicillin	Amoxil	*Commonly used to treat otitis media, sinusitis, and streptococcal infections*	BID or daily	Oral
	Amoxicillin + Clavulanate potassium	Augmentin	*idem*	BID or daily	Oral
Antibiotic (Cephalosporin)	Cefalexin	Keflex	Gram-positive infections	Daily	Oral
	Cefadroxil	Duricef	*Idem*	Daily	Oral

Antibiotic with Anti-inflammatory Effect (Macrolide)	Azithromycin	Zithromax	Streptococcal and Mycoplasma infections, syphilis, respiratory infections, etc. Also used as an anti-inflammatory in some diseases.	Daily	Oral
Non-steroidal Anti-inflammatory Drugs (NSAID)	Ibuprofen Naproxen Indomethacin Sulindac Aspirin	Advil Aleve Indocin Clinoril Aspirin	Anti-inflammatory	Daily	Oral
Corticosteroids (oral)	Prednisone	Prednisone	Anti-inflammatory	Daily	Oral
	Dexamethasone	Decadron	Anti-inflammatory	Daily	Oral
Corticosteroids (IV)	Methylprednisolone	Solumedrol	Anti-inflammatory	Short course over 1-3 days	IV
Antibody	Rituximab	Rituxan	*Autoimmune disease.*	One round over a day	IV
	Intravenous Immunoglobulin (IVIG)	Various	Inflammatory diseases and immunodeficiencies	Short course over 1-3 days	IV
Disease-modifying anti-rheumatic drug (DMARD)	Hydroxychloroquine	Plaquenil	Anti-rheumatic	Daily	Oral
	Methotrexate	Various	Anti-rheumatic	Daily	Oral or SC
	Mycophenolate mofetil	Cellcept	Anti-rheumatic	Daily	Oral

After keywords were highlighted, clinical records were reviewed in detail to ensure that the patient was taking the medication as listed. Abbreviations: Freq: Frequency of administration; Admin: Route of administration; IM: intramuscular; IV: intravenous; SC: subcutaneous.; BID (*bis in die*): two times a day; QID (*quater in die*) four times a day.

For medications taken daily, we determined the start and stop dates. For a short course of drugs administered over 1-3 days, we determined the start date only. In some cases, determining start and stop dates is challenging. For example, a patient/parent may have decided to discontinue a medication between two clinic visits but failed to recall the exact stop date. In these cases, we estimated the stop date using one of two methods: a) if the provider estimated a unit of time during which patient stopped taking the

medication, we used the midpoint of that unit of time (e.g. "patient stopped NSAIDs March 2017" would be coded as March 15, 2017; early March will be coded as March 1; late March will be coded as March 30; two weeks ago will be 14 days before the encounter date); b) if no estimate was given, we used the clinic visit date on which the provider reported the patient stopped taking the medication as the stop date. All stop dates for active medications were set at the last visit dates as there is no clue if the patient continues the medications or not. If a patient suspended the drug for less than a week, we would consider it to be a continuous use; otherwise, we would state two separate periods.

Patient's outcomes

The Stanford PANS clinic collects electronic patient questionnaires that caregivers fill out before each clinic visit. The questionnaire queries symptom-specific scales corresponding to the severity of the patient symptoms. We assessed two main scales: 1) the Global Impairment Scale[29], a scale ranging from 0 to 100, where the highest value indicates severe challenges for the patient to interact with others and carry on their daily activities, and the lowest value indicates a regular child without psychiatric problems. And 2) the Caregiver Burden Inventory[30], a scale ranging from 0 to 96, where the highest value indicates a severe difficulty for the patient's caregiver.

The *Medal* Algorithm

In this manuscript, we propose medication alignment for patient similarity (*Medal*) algorithm, that adapts a protein sequence alignment paradigm[34,35] to medication usage history. The alignment edit distance is used to estimate medication usage similarity, in order to construct a hierarchical clustering across the cohort. The algorithm is as follows: a) encode the medication history in a sequence representation; b) perform alignment of pairs of medication sequences; and c) compute a weighted patient pairwise edit distance.

Encode the medication history. For every pair of patients (e.g. patient 1 and 2), and one medication (e.g. prednisone), the usage was encoded into a vector space where every element of the vector corresponds to the usage status for one day. Thus, the dimension of the vector is determined by the first day of medication usage for both pairs of patients, and the last day for both patients. We considered missing values to be not missing at random (NMAR), and therefore assigned an explicit missing value. For example, consider two patients who are taking two rounds of a corticosteroid burst (3 days each, represented with letter 'C'), followed by two resting days (represented by the value '∅'). The first patient (represented as "CCC∅∅CCC∅∅") starts taking the corticosteroid on day 30 relative to the first day of diagnosis, while the second patient (represented as "∅∅CCC∅∅CCC") starts taking clindamycin on day 32. Therefore, in this illustrative example, the vectors would both have a length of 10, with an offset of two days (six days of clindamycin dosage, with two days of resting period). In real-world data, the medication patterns are far larger and more complex than those presented here. As a result, inspecting the differences between pairs of patients is not a straightforward task.

Sequence alignment. We used a dynamic programming approach to align the vectors generated in the previous step. First, an alignment matrix is generated, with the sequence for the one patient as column names, and the sequence for the other patient as row names. The matrix is initialized with zeros in the first column and row. Then, the matrix is filled with values considering at each cell only the three neighbors of that cell (left, upper-left diagonal, and top cell), starting from the top-left of the matrix, to the bottom-right.

Medal considers four medication usage patterns to fill values at each cell: start of medication, continuing medication, end of medication, and continuing gap. For any pair of patients, there are 16 possible combinations for how the medication is being administered. Four rules are used to assign values in each cell of the matrix: 1) if both

patients have the same medication usage pattern, then the minimum neighboring value is assigned; 2) if one of the patients has switching pattern (e.g. the first patient starting medication, and second patient continuing medication), then the maximum neighboring value is assigned; 3) if both patients have complete opposing patterns (e.g. first patient is starting medication while the second patient is ending medication), then the maximum neighboring value plus one is assigned; and 4) lastly, if one of the patients has a pattern involving missing values (e.g. first patient is starting medication, while the second patient is not taking anything), then the minimum neighboring value plus one is assigned.

Weighted edit distance. The edit distance is a measurement of the minimum number of operations needed to align two sequences. Traditionally in genome sequences, these operations could be insertion of a gap, deletion of a position, or substitution for an equivalent letter. However, for medication alignment, we only allow insertions, since we do not want to delete medication histories, and a substitution for an equivalent letter is difficult to assess. In Figure 6, the minimum distance, for the toy example above, is shown as the green path. The path starts from the bottom-right cell and ends at the top-left cell of the matrix. Only two insertions were needed to align the sequences. The purple path is an alternative sub-optimal alignment that required 3 insertions, but this distance is not selected.

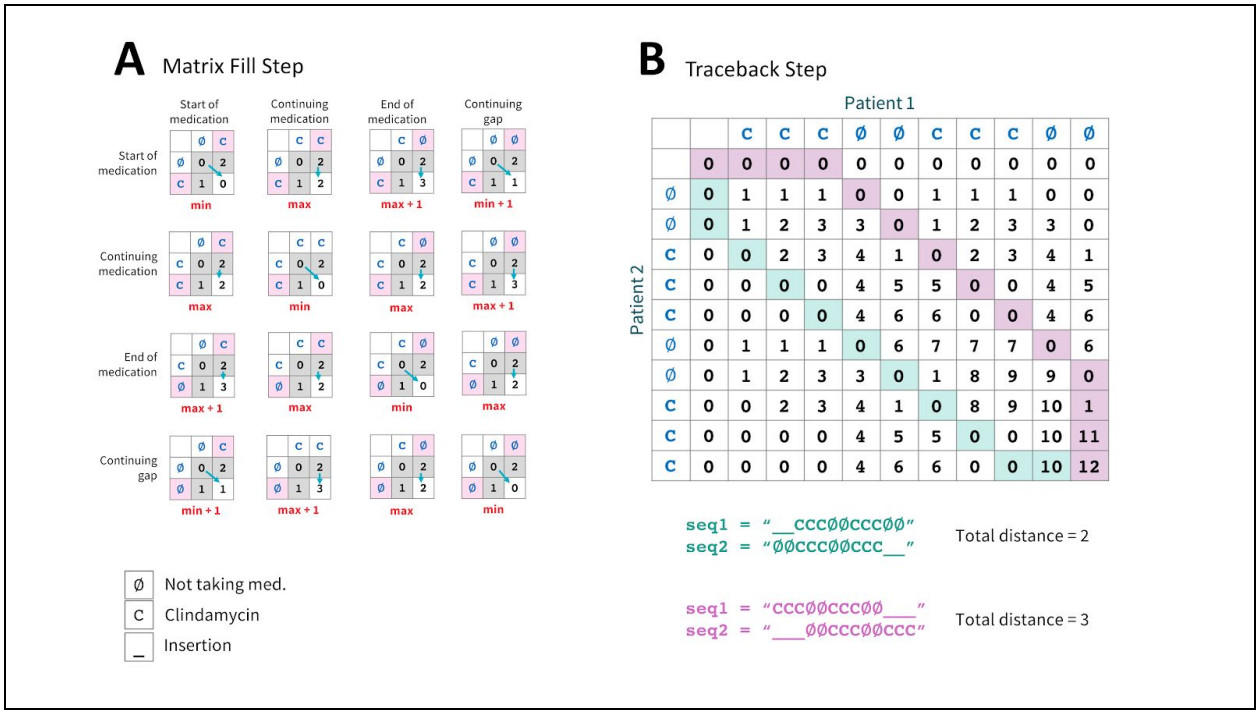

Figure 6. Dynamic programming approach to medication alignment. A) matrix fill step; B) Traceback step.

In order to estimate a similarity score between pairs of patients, we calculated a weighted sum of the individual medication edit distances (The weights are given by the size of the alignment in each medication, divided by the sum of alignment sizes). If a given medication was only being used by one of the patients, then the edit distance (as well as the size of alignment) would be equal to the total amount of usage days. If both patients were taking the medication, then the previously described dynamic programming approach would determine the edit distance. Table 4 shows an example of the composite score, indicating in days the amount of medication usage 'M', and non-usage '∅'; as well as the edit distance and alignment size.

Table 4. Example of medication alignment edit distances and sizes.

	Patient 1		Patient 2		Alignment	
Medication (M)	**∅**	**M**	**∅**	**M**	**distance**	**size**
Penicillin	768	113	0	0	881	881
Cephalosporin	709	1,369	2,067	11	2,075	4,153
Macrolide	1,382	129	1,504	7	1,328	2,839
NSAID	795	943	1,709	29	943	2,681
Cortisone	720	24	738	6	666	1,410
Antibody	2,610	56	2,664	2	2,663	5,329
DMARD	226	14	0	0	240	240
Weighted distance:					**1,761.24**	

Finally, a hierarchical clustering was performed using complete linkage. Hierarchical clustering is a popular technique that creates clusters with an ordering (hierarchy). The family of hierarchical clustering algorithms is very broad, for this work we selected an agglomerative (bottom-up) algorithm. This method starts by assigning each sample to its own cluster. Then, we use the similarity score (distance) calculated from our *Medal* algorithm to join the two most similar patients into a larger cluster. The algorithm continues to recursively aggregate clusters until all samples have been added to a single cluster. Distance between clusters are calculated using two linkage methods: a) complete, which calculates the distance between the two furthest points in both clusters; and b) average, which calculates the average distance between each point in one cluster to every point in the other cluster.

Statistical Evaluation

Our approach to patient similarity measurement was evaluated for space and time complexity. The change of representation from genomic sequences to daily intake of medications, as well as the change of operations needed to deal with the new type of data, should not affect the complexity of the dynamic programming approach. We provide summary statistics for space and time complexity needed in the PANS cohort.

Three quantitative methods were used to estimate the optimal number of clusters (*k*) in the PANS cohort: a) elbow method, b) Silhouette method, and c) gap statistic.

Elbow method. Measures the compactness of clusters by summing the within-cluster sum of squares. This method creates a plot of different *k* values (e.g. 1 to 10) and their corresponding total cluster variation. The selection of *k* is done by visual inspection of the plot, whenever an 'elbow' is found (a shift in the trend).

Silhouette method[27]. Measures the cohesion of clusters, or a metric of how well a patient belongs to a cluster. In this method, an average distance between each element and the rest of the elements in that cluster is compared to the average distance to neighboring clusters. The selection of *k* is done by choosing the maximum Silhouette value in a greedy forward selection (allowing a window of one step look-ahead).

Gap statistic[28]. Estimates the statistical comparison between the total intra-cluster variation and the null hypothesis without cluster assignment. The optimal number of clusters is chosen by the global maximum squared error value.

To compare trends between clusters for the Global Impairment score and Caregiver Burden Inventory, we conducted an analysis of variance (ANOVA) accounting for longitudinal sampling from each individual. Slopes were fit stratified by cluster using a linear mixed model to determine trends within clusters for both scores.

Acknowledgements

Competing interests

ALP declares that the research presented in this study was done while he was employed by Stanford University, but at the time of submission he is now employed by Genentech, Inc., a member of the Roche group.

The rest of the authors declare no competing interests.

Contributions

ALP designed the study. JF designed the patient inclusion/exclusion criteria. CMCL, AC, and JF reviewed medical notes. GLW, AP, AI, and ALP performed analysis of data. CDB, GLW, and JF provided interpretation of the results. ALP drafted the manuscript, and all authors contributed critically, read, revised and approved the final version.

Funding

Research reported in this publication was partially supported via institutional funds from Stanford University. CDB is a Chan Zuckerberg Biohub investigator. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Ethics approval

This study was approved by Stanford University's Institutional Review Board (IRB). Written informed consents were obtained from parents and assents were obtained from subjects aged 7 to 17 (all patients in this cohort) who were able to understand it (IRB protocol #26922). We did not have the opportunity to re-consent two patients at the time of data abstraction and analysis. The inclusion of their data is covered by a retrospective chart review protocol (IRB protocol #28533). Authors CMCL, AC, and JF served as honest brokers securing identifiable information. The remaining authors only had access to de-identified information. Non-human subject determination was provided (IRB protocol #46979).

Data Availability

The data that supports the findings of this study is available for research purposes upon written request, which will be reviewed on a case-by-case basis by Dr. Jennifer

Frankovich, director of the Stanford PANS clinic (http://med.stanford.edu/pans), and Stanford's Institutional Review Board (IRB).

References

1. Brown, S.-A. Patient Similarity: Emerging Concepts in Systems and Precision Medicine. *Front. Physiol.* **7,** 19 (2016).
2. Sharafoddini, A., Dubin, J. A. & Lee, J. Patient Similarity in Prediction Models Based on Health Data: A Scoping Review. *JMIR Med Inform* **5,** e7 (2017).
3. Ng, K., Sun, J., Hu, J. & Wang, F. Personalized Predictive Modeling and Risk Factor Identification using Patient Similarity. *AMIA Jt Summits Transl Sci Proc* **2015,** 132–136 (2015).
4. Panahiazar, M., Taslimitehrani, V., Pereira, N. L. & Pathak, J. Using EHRs for Heart Failure Therapy Recommendation Using Multidimensional Patient Similarity Analytics. *Stud Health Technol Inform* **210,** 369–373 (2015).
5. Wang, F. Adaptive semi-supervised recursive tree partitioning: The ART towards large scale patient indexing in personalized healthcare. *Journal of Biomedical Informatics* **55,** 41–54 (2015).
6. Lee, J. Patient-Specific Predictive Modeling Using Random Forests: An Observational Study for the Critically Ill. *JMIR Med Inform* **5,** e3 (2017).
7. Girardi, D. *et al.* Using concept hierarchies to improve calculation of patient similarity. *Journal of Biomedical Informatics* **63,** 66–73 (2016).
8. Li, L. *et al.* Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci Transl Med* **7,** 311ra174–311ra174 (2015).
9. Cahan, A. & Cimino, J. J. Visual assessment of the similarity between a patient and trial population: Is This Clinical Trial Applicable to My Patient? *Appl Clin Inform* **7,** 477–488 (2016).

10. Lee, J. *et al.* Privacy-Preserving Patient Similarity Learning in a Federated Environment: Development and Analysis. *JMIR Med Inform* **6,** e20 (2018).
11. Zhang, P., Wang, F., Hu, J. & Sorrentino, R. Towards personalized medicine: leveraging patient similarity and drug similarity analytics. *AMIA Jt Summits Transl Sci Proc* **2014,** 132–136 (2014).
12. Dai, Y., Lokhandwala, S., Long, W., Mark, R. & Lehman, L.-W. H. Phenotyping Hypotensive Patients in Critical Care Using Hospital Discharge Summaries. *IEEE EMBS Int Conf Biomed Health Inform* **2017,** 401–404 (2017).
13. Masnoon, N., Shakib, S., Kalisch-Ellett, L. & Caughey, G. E. What is polypharmacy? A systematic review of definitions. *BMC Geriatr* **17,** 230 (2017).
14. Shah, B. M. & Hajjar, E. R. Polypharmacy, adverse drug reactions, and geriatric syndromes. *Clin. Geriatr. Med.* **28,** 173–186 (2012).
15. Morandi, A. *et al.* Predictors of rehospitalization among elderly patients admitted to a rehabilitation hospital: the role of polypharmacy, functional status, and length of stay. *J Am Med Dir Assoc* **14,** 761–767 (2013).
16. Díaz-Caneja, C. M., Espliego, A., Parellada, M., Arango, C. & Moreno, C. Polypharmacy with antidepressants in children and adolescents. *Int. J. Neuropsychopharmacol.* **17,** 1063–1082 (2014).
17. Marengo, M. F. *et al.* Measuring therapeutic adherence in systemic lupus erythematosus with electronic monitoring. *Lupus* **21,** 1158–1165 (2012).
18. Justice, A. C. *et al.* Nonantiretroviral polypharmacy and adverse health outcomes among HIV-infected and uninfected individuals. *AIDS* **32,** 739–749 (2018).
19. Alshamrani, M., Almalki, A., Qureshi, M., Yusuf, O. & Ismail, S. Polypharmacy and Medication-Related Problems in Hemodialysis Patients: A Call for Deprescribing. *Pharmacy (Basel)* **6,** (2018).

20. LeBlanc, T. W., McNeil, M. J., Kamal, A. H., Currow, D. C. & Abernethy, A. P. Polypharmacy in patients with advanced cancer and the role of medication discontinuation. *Lancet Oncol.* **16,** e333–41 (2015).
21. Swedo, S. E. *et al.* Clinical presentation of pediatric autoimmune neuropsychiatric disorders associated with streptococcal infections in research and community settings. *J Child Adolesc Psychopharmacol* **25,** 26–30 (2015).
22. Brown, K. *et al.* Pediatric Acute-Onset Neuropsychiatric Syndrome Response to Oral Corticosteroid Bursts: An Observational Study of Patients in an Academic Community-Based PANS Clinic. *J Child Adolesc Psychopharmacol* **27,** 629–639 (2017).
23. Orlovska, S. *et al.* Association of Streptococcal Throat Infection With Mental Disorders: Testing Key Aspects of the PANDAS Hypothesis in a Nationwide Study. *JAMA Psychiatry* **74,** 740–746 (2017).
24. Pérez-Vigil, A. *et al.* The link between autoimmune diseases and obsessive-compulsive and tic disorders: A systematic review. *Neurosci Biobehav Rev* **71,** 542–562 (2016).
25. Frankovich, J. *et al.* Multidisciplinary clinic dedicated to treating youth with pediatric acute-onset neuropsychiatric syndrome: presenting characteristics of the first 47 consecutive patients. *J Child Adolesc Psychopharmacol* **25,** 38–47 (2015).
26. Swedo, S. E., Frankovich, J. & Murphy, T. K. Overview of Treatment of Pediatric Acute-Onset Neuropsychiatric Syndrome. *J Child Adolesc Psychopharmacol* **27,** 562–565 (2017).
27. Rousseeuw, P. J.1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 53–65 (1987). doi:10.1016/0377-0427(87)90125-7

28. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63,** 411–423 (2001).
29. Leibold, C., Thienemann, M., Farhadian, B., Willett, T. & Frankovich, J. Psychometric Properties of the Pediatric Acute-Onset Neuropsychiatric Syndrome Global Impairment Score in Children and Adolescents with Pediatric Acute-Onset Neuropsychiatric Syndrome. *J Child Adolesc Psychopharmacol* **29,** 41–49 (2019).
30. Farmer, C. *et al.* Psychometric Evaluation of the Caregiver Burden Inventory in Children and Adolescents With PANS. *J Pediatr Psychol* **43,** 749–757 (2018).
31. Elm, von, E. *et al.* The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* **370,** 1453–1457 (2007).
32. Mitra, S., Bastos, C. P., Bates, K., Pereira, G. S. & Bult-Ito, A. Ovarian Sex Hormones Modulate Compulsive, Affective and Cognitive Functions in A Non-Induced Mouse Model of Obsessive-Compulsive Disorder. *Front Behav Neurosci* **10,** 215 (2016).
33. Frankovich, J., Longhurst, C. A. & Sutherland, S. M. Evidence-based medicine in the EMR era. *N Engl J Med* **365,** 1758–1759 (2011).
34. Lipman, D. J. & Pearson, W. R. Rapid and sensitive protein similarity searches. *Science* **227,** 1435–1441 (1985).
35. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *PNAS* **85,** 2444–2448 (1988).