Deep clustering to identify subgroups of multivariate trajectories in longitudinal biomedical datasets

Bhargav Vemuri

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Peter Tarczy-Hornoch, Chair

Jennifer Hadlock

Susan Shortreed

Program Authorized to Offer Degree:

Biomedical Informatics and Medical Education

University of Washington

Abstract

Deep clustering to identify subgroups of multivariate trajectories in longitudinal biomedical datasets

Bhargav Vemuri

Chair of the Supervisory Committee:

Peter Tarczy-Hornoch

Department of Biomedical Informatics and Medical Education

Unsupervised patient subgrouping in longitudinal biomedical datasets enables the discovery of distinct temporal phenotypes that capture heterogeneity in disease progression, treatment response dynamics, developmental trajectories, telemonitoring, and more. One-stage multivariate time series (MVTS) deep clustering methods are well-suited to this task because they (1) jointly model multiple longitudinal variables and (2) integrate missing data imputation, representation learning, and clustering into a unified framework. Recent state-of-the-art MVTS deep clustering approaches include Variational Deep Embedding with Recurrence (VaDER; de Jong et al., 2019) and Clustering Representation Learning on Incomplete time-series data (CRLI; Ma et al., 2021). In this work, we apply CRLI in two real-world longitudinal biomedical contexts and evaluate its performance against VaDER using 20 synthetic MVTS datasets of our own design. Our overarching question was: how and when are one-stage MVTS clustering methods (VaDER, CRLI) useful in biomedical research data exploration?

In Aim 1 (*Assessing the ability of CRLI to detect meaningful trajectories in a sparse, irregular, biased real-world dataset*), we explored CRLI's capacity to detect multivariate trajectories in the electronic health record (EHR). Temporal EHR data is marred by irregular measurement intervals, high missingness, and multiple biases (selection, measurement, time-related). We assessed how well CRLI handles these hurdles in the context of identifying GLP-1 medication (semaglutide, dulaglutide, etc.) treatment response subgroups in the NIH *All of Us* Research Study. We showed that (1) CRLI can be used to identify post-treatment multivariate response trajectories in the EHR and (2) this is possible despite a small cohort (n=336) and infrequent measurements.

In Aim 2 (*Assessing the ability of CRLI to detect meaningful trajectories in a high-dimensional, multimodal, prospective dataset*), we applied CRLI to another real-world data source, the Adolescent Brain Cognitive Development (ABCD) Study, a longitudinal observational cohort with a prespecified assessment protocol, including a consistent follow-up schedule and a high retention rate (98.9%). This dataset allowed us to explore physical health trajectories (pubertal hormones, anthropometrics) as we did in Aim 1, but also mental health trajectories, as measured by 8 Child Behavior Checklist (CBCL) syndrome scales. We calculated cluster associations with mental health outcomes to better characterize cluster differences. We showed that (1) given longitudinal and static variables, CRLI identified longitudinal trajectories that had non-uniform associations with static variables, providing a basis for testable clinical hypotheses, and (2) CRLI identified clusters that could not have been identified with a single timepoint or single variable alone.

In Aim 3 (*Assessing the ability of CRLI and VaDER to detect trajectories in synthetic datasets under diverse data constraints*), we designed a framework using the mockseries Python

package that let us rapidly generate unique MVTS datasets by sampling from a range of values for various datasets characteristics (time series length, noise, missingness, number of clusters, number of samples). We also incorporated the ability to modify time series variable properties (trend, rate of change, seasonality) by designing 5 distinct variable styles inspired by biomedical trends we observed in Aims 1 and 2 and the literature. We reported VaDER and CRLI performance on 4 external clustering validation indices (purity, RI, ARI, NMI) across 20 synthetic datasets. We showed that (1) practitioners should be wary of novel methods that do not report performance on adjusted metrics (ARI, AMI), (2) 2D visualizations are an invaluable interpretability tool, especially when there are too many longitudinal variables to understand on an individual basis, and (3) while CRLI generally outperforms VaDER, neither method achieved across-the-board ARI dominance.

Cross-cutting contributions that emerged across the aims were as follows: (1) we observed that internal clustering validation indices (Calinski-Harabasz, Silhouette, Davies-Bouldin, S_Dbw validity) were rarely concordant, making the selection of optimal cluster number in Aims 1 and 2 complicated, (2) cohort selection criteria that required a minimum number of repeat measurements across multiple longitudinal variables resulted in final cohorts that may not have generalized well to the population and/or an external validation dataset, (3) method performance in Aim 3 as measured by Adjusted Rand Index (external clustering validation index) was subpar compared to other indices that have been reported in the literature, casting doubt on trustworthiness of clusters identified in previous Aims, and (4) visual (qualitative) inspection and interpretation of identified clusters is a necessary complement to quantitative clustering result evaluation (by internal and/or external clustering validation index) for a holistic understanding of trajectory differences between clusters.

Acknowledgements

First and foremost, I would like to acknowledge my parents, Vasanta and Ranga Vemuri. As my very first teachers, you always emphasized the sanctity of scholarship and fostered the scientific curiosity which enabled me to complete this dissertation. Your tireless and uncompromising dedication to my growth is the reason I am where I am today. My sincerest thanks to my younger sister, Sarvani, for being the embodiment of discipline and achievement and a role model I can look up to. I am immensely grateful for your ever presence in my life and could not have asked for a stronger family unit.

I want to express my gratitude to the Boyce, Mangipudi, and Putrevu families for making Seattle feel like a home away from home when I needed it most. Your hospitality and warmth were instrumental in getting me from start to finish.

To my first advisor at UW and committee chair, Peter Tarczy-Hornoch, I am deeply grateful for your steadfast support, sagely advise, timely encouragement, and firm guidance. To my research advisor and co-chair, Jennifer Hadlock, I thank you for your infinite patience with, unwavering belief in, and unreserved support of me and my research ideas. Thank you also to my other committee members, Susan Shortreed and Abie Flaxman, for your insights, motivation, and willingness to be a part of this undertaking.

To my fellow 2021 BIME cohort mates (Ehsan, Ashmi, Namu, Kevin, Sharon, Serena, Xinyang, Yile, and others), I appreciate your camaraderie and companionship on this adventure. I'd like to thank the many others in the BIME community (Marni Levy, Melissa Espinoza, Shayla Simuel, Heather Clausnitzer, Heidi Krueger, Moiya Callahan, Dina Dwyer, Kathryn Hagy, Nick Reid, Oliver J. Bear Don't Walk IV, Chethan Jujjavarapu, Hannah Burkhardt, to name a few) for your

assistance and guidance at various points over the past several years. I would also like to extend my thanks to members of the Hadlock Lab and ISB (Yeon Mi Hwang, Qi Wei, Andrew Baumgartner, Alex Ralevski, Sevda Molani, Thea Swanson, Noa Mardiks Rappaport, Gwênlyn Glusman, Connor Kelly, Evan Pepper, and others) for the same.

Thank you to WRF (Kim Emmons, Loretta Little, Meher Antia), ITHS and the TL1 program (Hilaire Thompson, Megan Moore, Sara Teklehaimanot, Russ Lackey), and IMDS (Sean Mooney, Shachi Mittal, Neha Sathe, Ti Haynes, Nadege Mohr) for your support, expertise, and opportunities to grow outside of the lab.

Lastly, thank you to my friends, family, peers, colleagues, and anyone else who made this work possible. The board game nights, hikes, bouldering sessions, potlucks, concerts, flag football games, cabin trips, skiing expeditions, escape rooms, early mornings, and late nights spent with you kept me going. I look forward to making more memories with you all in the future.

This work was supported in part by the Institute of Translational Health Sciences under grant TL1 TR002318 and by the National Institutes of Health, National Library of Medicine (NLM) University of Washington Biomedical Informatics and Data Science Research Training Program (Grant Nr LM 007442). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Table of Contents

Glossary

EHR = electronic health record
All of Us = NIH *All of Us* Research Program
ABCD = Adolescent Brain Cognitive Development

MVTS = multivariate time series
IMVTS = incomplete multivariate time series

DC = deep clustering
DTSC = deep time series clustering
VaDER = Variational Deep Embedding with Recurrence
CRLI = Clustering Representation Learning on Incomplete time-series data
CVI = clustering validation index

UCR = University of California, Riverside
UCI = University of California, Irvine
UEA = University of East Anglia

KS = k-Shape
DEC = Deep Embedded Clustering
IDEC = Improved Deep Embedded Clustering
DTC = Deep Temporal Clustering
DTCR = Deep Temporal Clustering Representation
VaDE = Variational Deep Embedding

MAR = Missing At Random
MNAR = Missing Not At Random
MCAR = Missing Completely At Random

1. Executive summary

1.1. Motivation

Unsupervised patient subgrouping in longitudinal biomedical datasets enables the discovery of distinct temporal phenotypes that can capture heterogeneity in disease progression, treatment response dynamics, developmental trajectories, telemonitoring, and more. Identifying such phenotypes can have implications for intervention design, treatment personalization, drug repurposing, and preventative care. Multivariate longitudinal approaches move beyond cross-sectional analyses to allow for richer characterization of subgroups across many measures simultaneously. Real-world (observational) data reflects true-to-life heterogeneity that may not be captured in carefully designed and controlled trials.

End-to-end (one-stage) multivariate time series (MVTS) deep clustering methods are particularly well-suited to this task because they (1) jointly model multiple longitudinal variables and (2) integrate missing data imputation, representation learning, and clustering into a unified framework. Recent state-of-the-art approaches include Variational Deep Embedding with Recurrence (VaDER; de Jong et al., 2019) and Clustering Representation Learning on Incomplete time-series data (CRLI; Ma et al., 2021). In this dissertation, we seek to characterize how and when such end-to-end MVTS deep clustering methods are useful in biomedical research data exploration. To that end, we apply CRLI in two real-world longitudinal biomedical contexts and evaluate its performance against VaDER using 20 synthetic MVTS datasets of our own design (Table 1). In Aim 1, we assess the ability of CRLI to detect meaningful trajectories in a sparse, irregular, biased real-world dataset (NIH All of Us EHR). In Aim 2, we assess the ability of CRLI to detect meaningful trajectories in a high-dimensional, multimodal, prospective dataset (The Adolescent Brain Cognitive Development Study). In Aim 3, we assess the ability of CRLI and VaDER to detect trajectories in synthetic datasets under diverse data constraints.

	Aim 1	Aim 2	Aim 3
Research area	Glucagon-like peptide-1 (GLP-1) receptor agonist (RA) treatment response	Physical & mental adolescent development	Benchmarking for MVTS clustering
Data source type	Electronic health record	Longitudinal observational cohort	Synthetic
Dataset	NIH *All of Us* Research Program	Adolescent Brain Cognitive Development (ABCD) Study	Newly-developed framework to generate random MVTS datasets
Tools	CRLI	CRLI	mockseries, VaDER, CRLI
MVTS clustering applicability	Multiple treatment outcome measures collected in routine clinical care	Many multidomain developmental measures assessed yearly	Range of dataset & time series properties, incl. # of variables, missingness rate, sample size, noise, time series length
Data challenges	Short time series lengths, irregular sampling rates, data sparsity, nonlinearity, variable inter-correlations		

Table 1. Aims overview

1.2. Related work

In terms of our biomedical areas of interest, deep representation learning has been applied to the temporal EHR to characterize diabetes progression trajectories, but EHR-derived insights specific to GLP-1 medications have been limited to (1) effectiveness and risks, (2) utilization, discontinuation, and reinitiation patterns, and (3) treatment responsiveness.[1–10] Researchers have used the ABCD dataset extensively to characterize mental health outcomes, like self-injurious behavior, with respect to prevalence, predictors, and transitions.[11–17] Physical health measures, like pubertal timing and and body weight, and their relation to psychopathology, have also been characterized in ABCD and other adolescent datasets.[18–22] However, trajectory modeling has been limited to univariate approaches or relied on model-based statistical approaches for multivariate analysis.[23–25]

1.3. Open questions and gaps in current knowledge

Deep time series clustering methods have been catalogued and benchmarked by several others.[26–29] But, thus far, the application of VaDER and CRLI to biomedical datasets has been limited, despite their demonstrated capabilities.[30–32] Furthermore, the head-to-head evaluation of

these two methods has lacked in diversity of MVTS datasets and rigor of validation metrics reported.[33] Thus, drawing from (a) our biomedical areas of interest (previous section), (b) commonly utilized sources of longitudinal observational data, and (3) the aforementioned evaluation inadequacies, we devised our three aims to better understand how and when one-stage MVTS clustering methods (VaDER, CRLI) are useful in biomedical research data exploration (Figure 1).

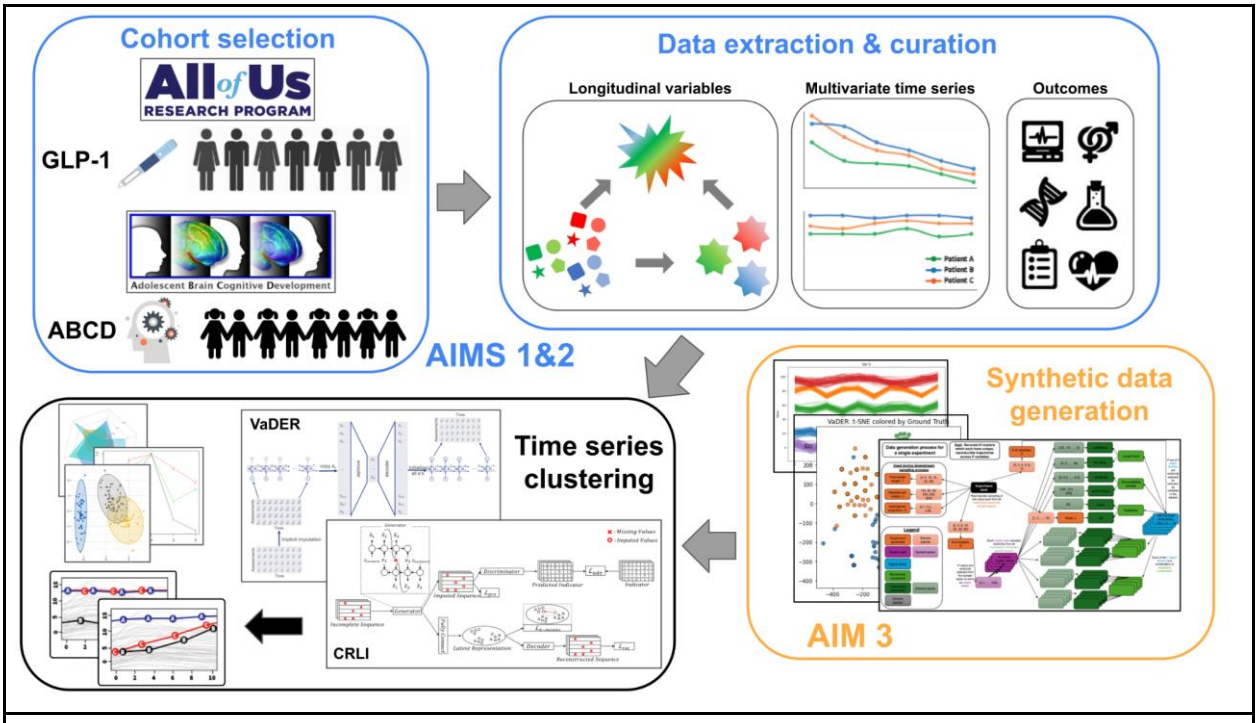


Figure 1. Graphical abstract of our work

1.4. Overview of chapters

In **Chapter 2**, we discuss in depth the background and prior studies relevant to all aims. This includes the structure of multivariate time series, types of longitudinal real-world data sources, data challenges encountered when working with such datasets, deep representation learning-based methods for clustering MVTS, including VaDER and CRLI in particular, and methods for evaluating clustering results.

In **Chapter 3 (Aim 1: Assessing the ability of CRLI to detect meaningful trajectories in a sparse, irregular, biased real-world dataset)**, we explore CRLI's capacity to detect multivariate trajectories in the electronic health record (EHR). Temporal EHR data is marred by irregular measurement intervals, high missingness, and multiple biases (selection, measurement, time-related). We assess how well CRLI handles these hurdles in the context of identifying GLP-1 medication (semaglutide, dulaglutide, etc.) treatment response subgroups in the NIH *All of Us* Research Study.

In **Chapter 4 (Aim 2: Assessing the ability of CRLI to detect meaningful trajectories in a high-dimensional, multimodal, prospective dataset)**, we apply CRLI to another real-world data source, the Adolescent Brain Cognitive Development (ABCD) Study, a longitudinal observational cohort with a prespecified assessment protocol, including a consistent follow-up schedule and a high retention rate (98.9%). This dataset allowed us to explore physical health trajectories (pubertal hormones, anthropometrics) as we did in Aim 1, but also mental health trajectories, as measured by 8 Child Behavior Checklist (CBCL) syndrome scales. We calculate cluster associations with mental health outcomes to better characterize cluster differences.

In **Chapter 5 (Aim 3: Assessing the ability of CRLI and VaDER to detect trajectories in synthetic datasets under diverse data constraints)**, we design a framework using the mockseries Python package that lets us rapidly generate unique MVTS datasets by sampling from a range of values for various datasets characteristics (time series length, noise, missingness, number of clusters, number of samples). We also incorporate the ability to modify time series variable properties (trend, rate of change, seasonality) by designing 5 distinct variable styles inspired by biomedical trends we observed in Aims 1 and 2 and the literature. We report VaDER and CRLI performance on 4 external clustering validation indices (purity, RI, ARI, NMI) across 20 synthetic datasets.

In **Chapter 6**, we summarize key findings, limitations, and future directions for each aim. We also discuss limitations that apply to all aims, cover recent advances in the field that have been made since completion of our work, and promising new areas that are under active investigation.

1.5. Contributions

In **Aim 1**, we explored CRLI's capacity to detect multivariate trajectories of chronic disease treatment response in the electronic health record (EHR). Specifically, in a cohort of 336 NIH *All of Us* participants taking GLP-1 medications for at least 2 years, who had measurements of HbA1c, serum creatinine, BMI, DBP, and SBP at least once in the 6 months prior to starting the medication and twice in the 2 years after starting the medication, we reported a 5-cluster result and a 2-cluster result. Both clustering results mapped to clinically relevant ranges across multiple variables, showing that CRLI was able to identify quantitatively and qualitatively distinct clusters. In both clustering results, we saw HbA1c decrease for all clusters, as expected. Interestingly, BMI was quite different between clusters, but relatively stable within each cluster. We generated mean lineplots with 95% confidence intervals for each cluster, allowing us to visualize cluster separation and lability. In the context of our overall question, we demonstrated that CRLI is useful for temporal EHR subtyping, even when data is sparse and irregularly sampled, though cluster validation can be complicated by conflicting indices in the absence of a qualitative interpretation (like clinical cogency as determined by a domain expert).

In **Aim 2**, we explored CRLI's capacity to detect multivariate trajectories of physical and mental health changes during adolescence in a prospective observational cohort. In a cohort of 2,923 female participants from the ABCD Study, we identified 3 multivariate trajectories of physical development across pubertal hormones (DHEA, testosterone, estradiol) and anthropometrics (BMI calculated from height and weight). In a cohort of 310 participants who exhibited self-

injurious behaviors (SIB) at the year 3 ABCD assessment, we identified multivariate trajectories of psychopathology (8 CBCL syndrome scales) in the years leading up to the SIB. We conducted association testing between cluster membership and outcomes at year 3 and year 4, revealing several significant associations with KSADS symptoms and diagnoses. CRLI was able to identify specific inflection points in adolescent development where a subgroup may have an important shift in rate of change of one measure or more. In particular, in Experiment 2 (SIB cohort), we saw more marked cluster differences in thematic areas like pleasure, worry, and depression, whereby some clusters had a "protective" influence against certain outcomes.

In **Aim 3**, we explored VaDER and CRLI's capacities to detect multivariate trajectories in several simulated datasets under diverse data conditions. Using Python package mockseries, we generated 20 MVTS datasets with varying lengths, missingness proportions, noise levels, cluster numbers, and sample sizes and evaluated performance of VaDER and CRLI against ground truth labels using four external validation metrics (purity, Normalized Mutual Information, Rand Index, Adjusted Rand Index). Our novel synthetic dataset generation approach complements the real-world data-based MVTS clustering method benchmarks that are generally (1) univariate, (2) non-biomedical, and (3) low missingness. We found that CRLI was equal to or outperformed VaDER on 19/20 datasets. However, ARI was by far the most conservative of the 4 validation indices and only crossed 0.5 in 3/40 evaluations (20 datasets*2 methods). As ARI was not reported by Ma et al. in their evaluation of CRLI, our results cast doubt on whether it is actually as performant as marketed in the original paper. Therefore, in the context of our overall question, we demonstrated that CRLI is more able to detect clusters than VaDER under diverse data constraints, but that (a) VaDER is possibly more performant if given enough training data and (b) neither method performs well when assessed by ARI.

Cross-cutting contributions that emerged across the aims were as follows: (1) we observed that internal clustering validation indices (Calinski-Harabasz, Silhouette, Davies-Bouldin, S_Dbw validity) were rarely concordant, making the selection of optimal cluster number in Aims 1 and 2 complicated, (2) cohort selection criteria that required a minimum number of repeat measurements across multiple longitudinal variables resulted in final cohorts that may not have generalized well to the population and/or an external validation dataset, (3) method performance in Aim 3 as measured by Adjusted Rand Index (external clustering validation index) was subpar compared to other indices that have been reported in the literature, casting doubt on trustworthiness of clusters identified in previous Aims, and (4) visual (qualitative) inspection and interpretation of identified clusters is a necessary complement to quantitative clustering result evaluation (by internal and/or external clustering validation index) for a holistic understanding of trajectory differences between clusters.

2. Background and prior studies

Applying clustering approaches to longitudinal biomedical datasets enables the discovery of distinct patient subgroups that reflect heterogeneous temporal patterns (Figure 2). Though the end products are referred to by various names in the literature (progression patterns, longitudinal phenotypes, clinical trajectories, trajectory subgroups, and so on), the underlying goal is the same: group individuals together that have similar trajectories of one or more repeated measures.[34,35] This approach can generate insights in areas like disease progression, treatment response dynamics, developmental trajectories, telemonitoring, and more. Specific clinical areas in which trajectory clustering approaches have already been deployed include cardiometabolic disease, neurodegenerative disease, adult-onset asthma, mental disorders, and hospital-related emergencies.[36]

In this chapter, we discuss longitudinal biomedical data sources, the complexities arising from their temporality, the deep learning approaches that have been devised to handle those complexities, and the specific methods that we employed in this work to characterize the value of deep MVTS clustering in biomedical research data exploration.

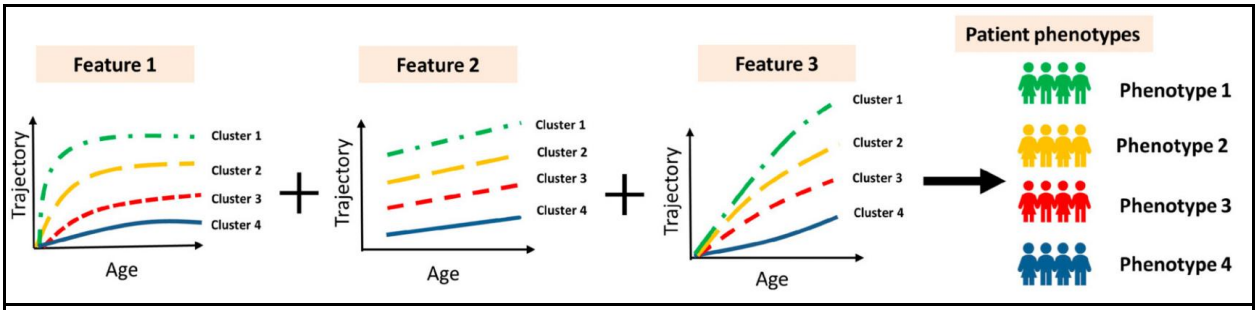


Figure 2. Integrative clustering of multiple longitudinal features[35]

2.1. Multivariate time series definition and analyses

A time series is a chronologically ordered sequence of V-variate data points recorded at specific time intervals. When $V = 1$, it is a univariate time series; otherwise, it is a multivariate time series

(MVTS). Figure 3 shows an MVTS dataset with 6 timepoints (T0-T5), 3 variables (A, B, C), and N samples stacked into a 3-dimensional array.[37] When the intervals between observations are not consistent or regularly spaced, the time series is considered to be irregularly sampled, leading to missingness (Figure 4). Time series imputation aims to fill in missing values with realistic values to facilitate downstream analysis (Figure 5).[38]

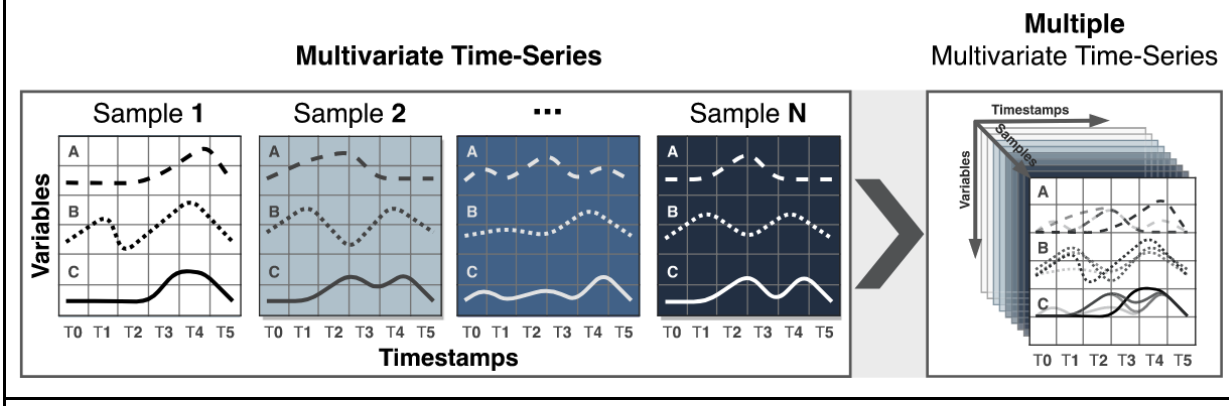


Figure 3. MVTS data as a 3D array (timestamps, variables, samples)[37]

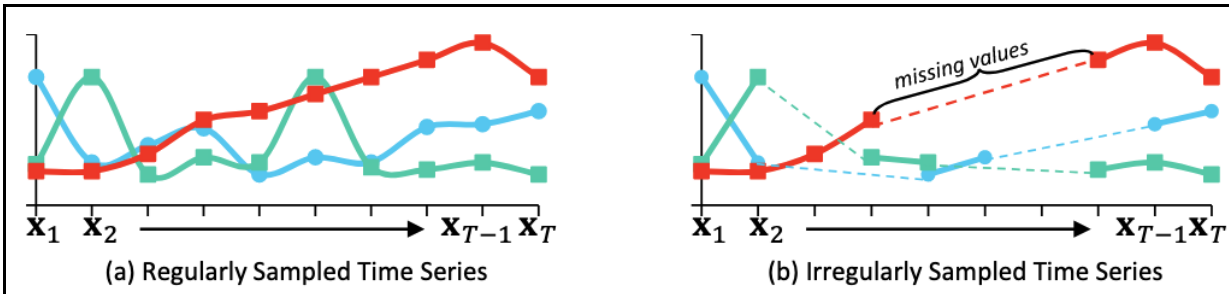


Figure 4. Regularly and irregularly sampled multivariate time series (V = 3)[38]

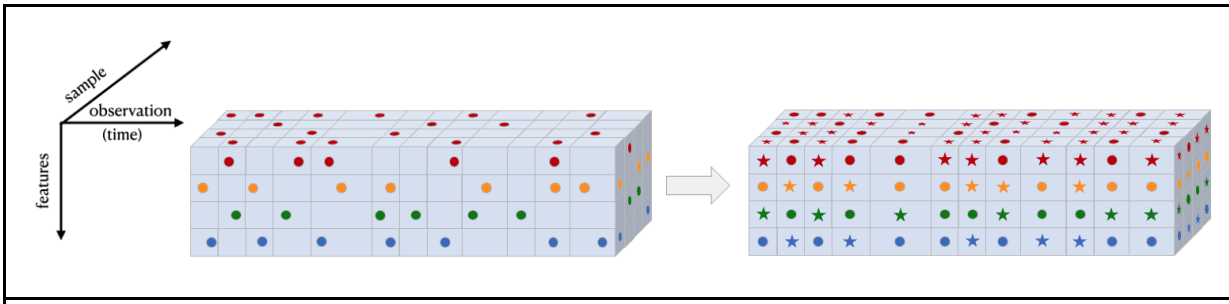


Figure 5. Before and after imputation of irregularly sampled MVTS[39]

Other time series analysis tasks include indexing, classification, forecasting, clustering, anomaly detection, regression, similarity search, segmentation, and more. Non-biomedical application domains for these tasks include finance, human activity and speech recognition, traffic flow, IoT sensors, and cyber-physical systems. Middlehurst et al. provide a flowchart (Figure 6) for selecting the appropriate task given the analytical goal. Our focus is on unsupervised time series clustering as a means to identify multivariate temporal phenotypes in longitudinal biomedical datasets; non-clustering tasks have been described elsewhere.[27,38,40,41]

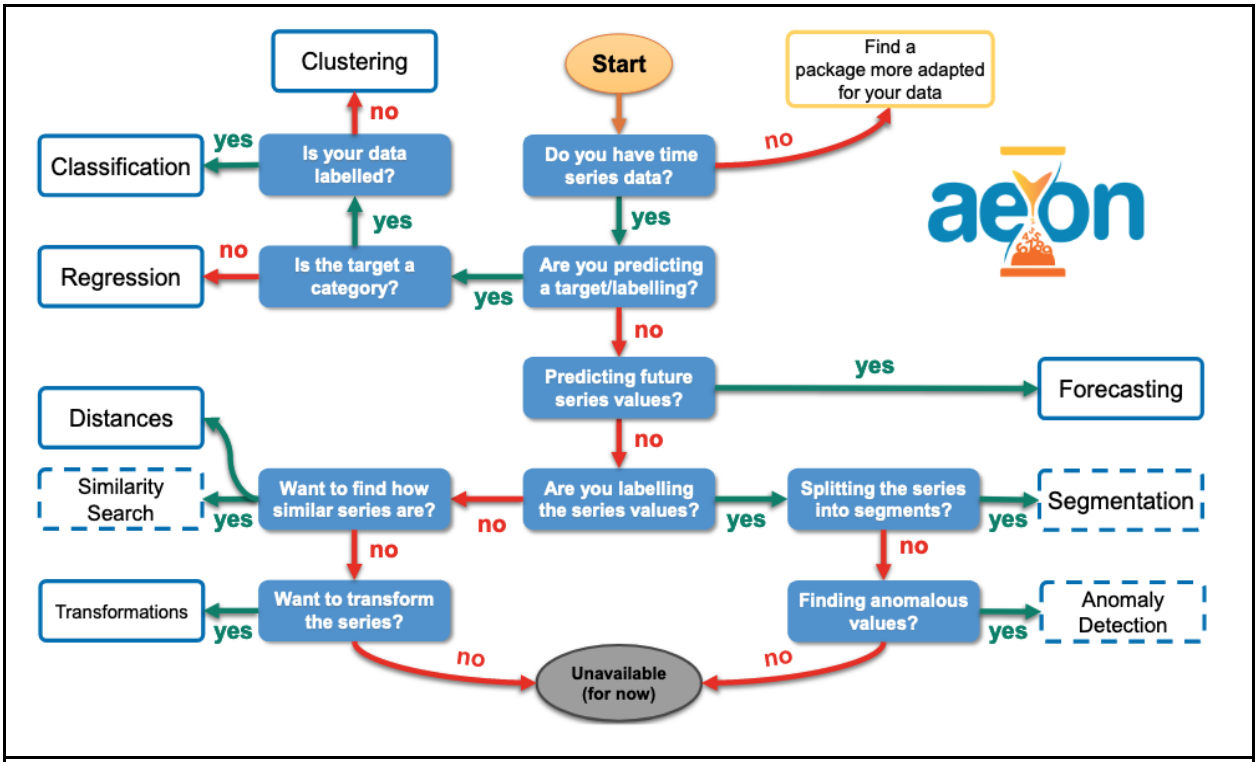


Figure 6. Flowchart for selecting an appropriate time series analysis task[41]

2.2. Longitudinal real-world data sources

Common sources of observational, or real-world, data (RWD) include electronic health records (EHR), administrative data, claims data, patient registries, patient-generated health data, and observational cohorts.[42–44] Each source has unique benefits and drawbacks with regard to

reflecting the temporal patient, or participant, experience (Table 2).[a] Since EHR data is collected during the course of routine interaction with the healthcare system, it is the most realistic representation of a patient's longitudinal medical status. EHRs can contain multiple data types, such as demographics, prescription history, laboratory tests, diagnoses, imaging, and unstructured clinical notes. However, the EHR was not designed as a research tool and therefore has no study design or guaranteed measurement regularity between patients (Figure 7).[46–48] On the other hand, prospective longitudinal cohort studies are restricted by study participation, and thus may be less representative of true population heterogeneity. But they benefit from having a prespecified study protocol and structured, regular assessment timepoints that results in more complete data availability overall. More specifics on longitudinal EHR and observational cohort data can be found in Chapters 3 and 4, respectively.

Characteristics	RWD-EHR	Prospective longitudinal cohort study or registry
Definition	Data from EHR relating to patient health status and/or the delivery of healthcare routinely collected from a variety of sources	Non-interventional clinical study, prospectively collecting data on a group of patients with a particular disease or symptom
Patient population	Broad, encompassing medical system or population area	Restricted by study participation
Data types	High dimensional	High no, limited by research design and variables for collection decided a priori
	Data collected as part of patient care from both patients and physicians	Structured data collection and questionnaires
Data presence	Sparce, noisy	Structured, same data collected on all participants
	Missingness not at random	Fairly complete
Scale	Large, thousands to millions	Modest, hundreds to thousands
Generalisability	Strong local structure can restrict generalisability Incorporating real-life noise into the analyses improves applicability to real life settings	Easily replicable in similar designed cohorts Generalisability restricted by patient selection and data not always directly implementable to real life settings.

Table 2. Characteristics of observational data sources[43]

[a] Regulatory bodies differ slightly on what they consider to be RWD (UK NICE vs. US FDA)[42,45]

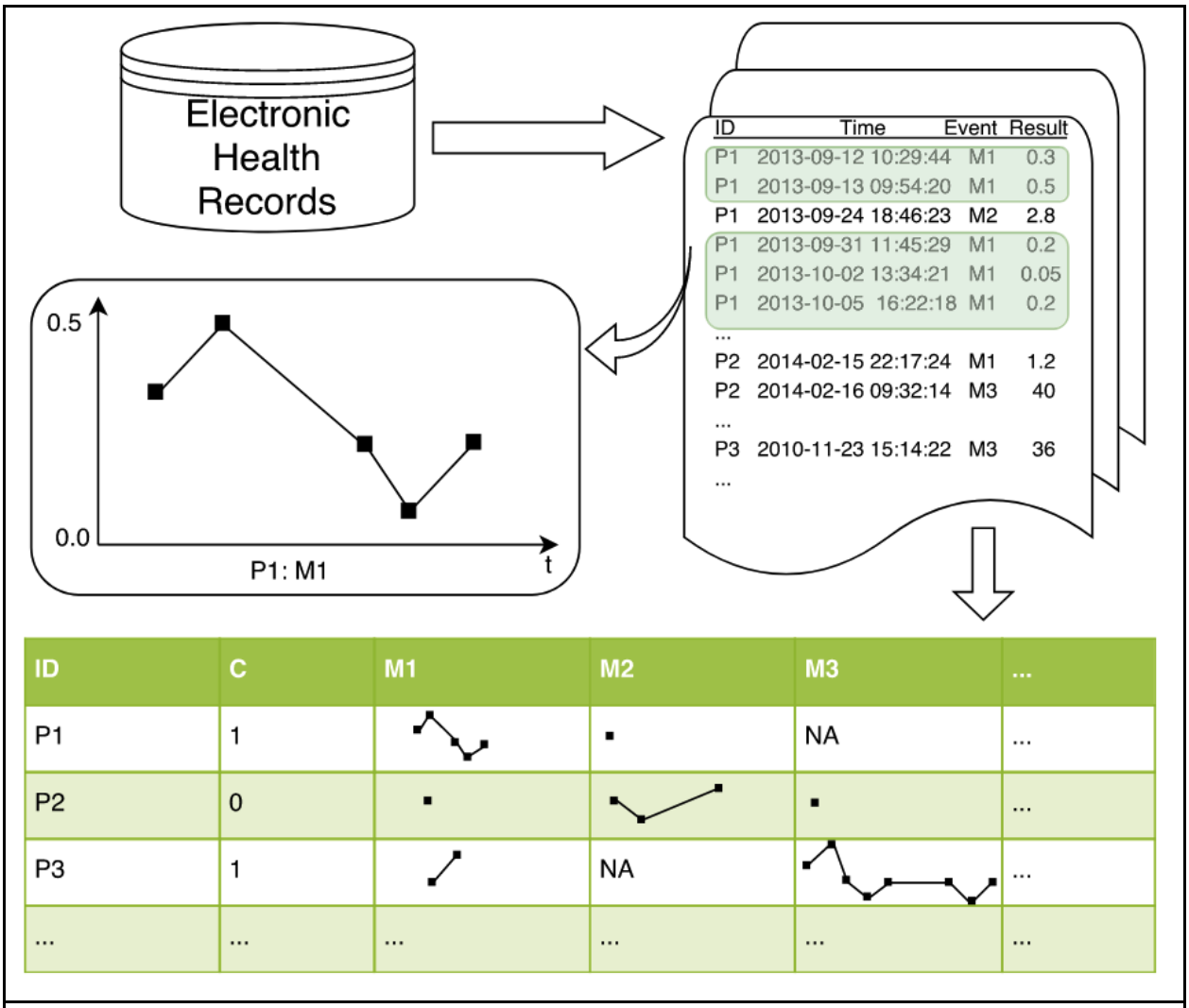


Figure 7. Inter-patient variability of longitudinal measurement regularity in the EHR[46]

2.3. Longitudinal data challenges

Across longitudinal data sources, there arise several data complexities unique to the temporal component that complicate time series analyses. These include missingness, risk of bias, and high dimensionality.

2.3.1. Missingness

2.3.1.1. Sampling rate, density, and length

Though most time series analysis models assume sequences are evenly spaced and complete, real-world data rarely adheres to that assumption. MVTS can suffer from intra- and inter-series

sampling irregularity (Figure 8). Intra-series irregularity occurs when time interval length within the same series is not uniform. This can be handled by selecting a fixed interval and treating timepoints without data as missing (Figure 9). Inter-series irregularity occurs when the sampling rate differs across variables. This can be handled by (1) increasing the granularity of the fixed interval to accommodate the higher sampled variable (MBP in Figure 8), thereby making for a sparser time series in the lower sampled variables (PH in Figure 8), or (2) reduce the length of the higher sampled variable by using a method like Piecewise Aggregate Approximation (see Section 3.4.3.1 for details). While longitudinal cluster analyses have historically been limited to few repeat measurements per subject (e.g., less than 10), densely sampled temporal datasets are becoming increasingly common.[49] Intensive longitudinal data (ILD) refers to time series with a high measurement frequency (20 or more repeated observations within a time interval) which can capture trends at a more granular level.[50,51]

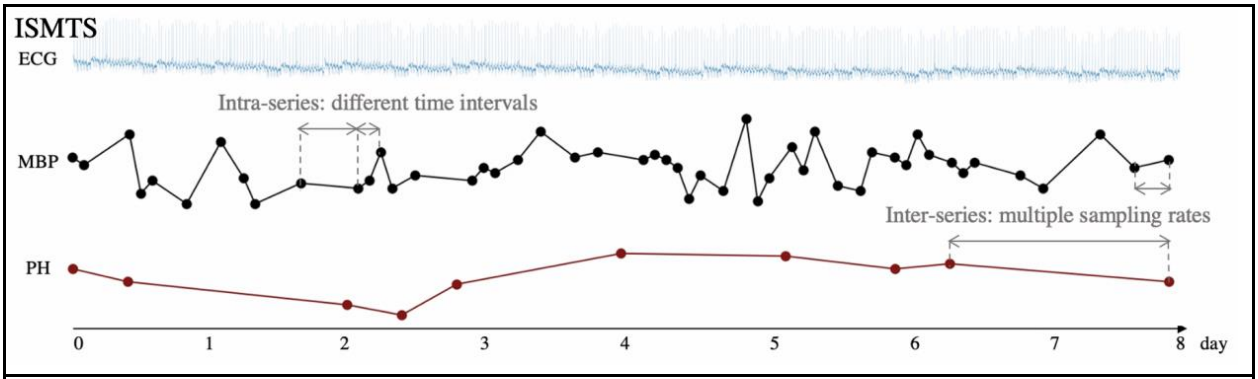


Figure 8. Intra- and inter-series sampling rate differences in the temporal EHR[52]

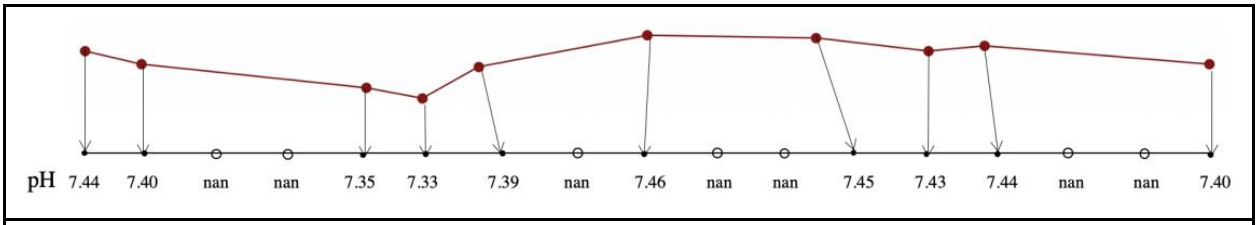


Figure 9. Conversion of irregularly sampled data to missingness[52]

2.3.1.2. Reasons for and patterns of missingness

Missingness in biomedical longitudinal datasets can occur for a number of reasons, depending on the type of data source and the overall study design. These include participant drop out, missed visits, faulty diagnostic equipment, time constraints, lack of insurance coverage, changing monitoring guidelines, and so on. Nagin et al. distinguished two forms of missing data in longitudinal studies: intermittent missingness and nonrandom attrition (i.e., dropout). Accordingly, Figure 10 shows unique assessment patterns for four hypothetical participants in a six-period study.[34] Lu et al. described four kinds of commonly seen longitudinal data structures based on variations in (1) spacing between data points and (2) data balance across individuals (Figure 11). Scenario 1 shows equal time spacing and balanced data, Scenario 2 shows equal time spacing and unbalanced data, Scenario 3 shows unequal time spacing and balanced data, and Scenario 4 shows unequal time spacing and unbalanced data.[53] Extending this to the multivariate context, if we consider each Scenario to be a separate variable in a dataset, we can see the previously discussed intra- and inter-series sampling irregularity across variables.

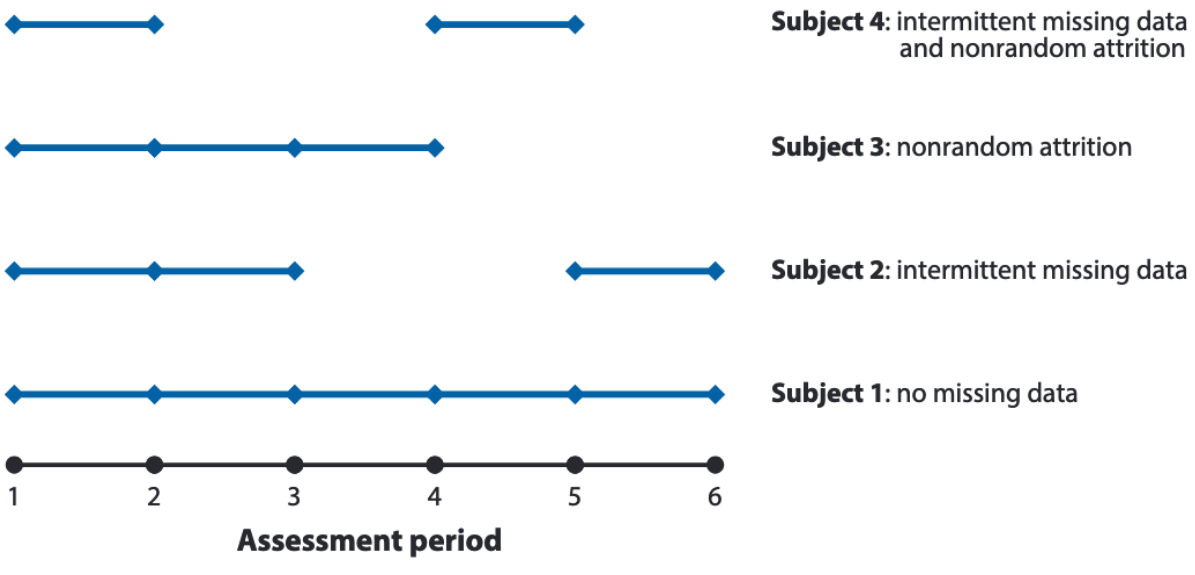


Figure 10. Intermittent missingness and nonrandom attrition[34]

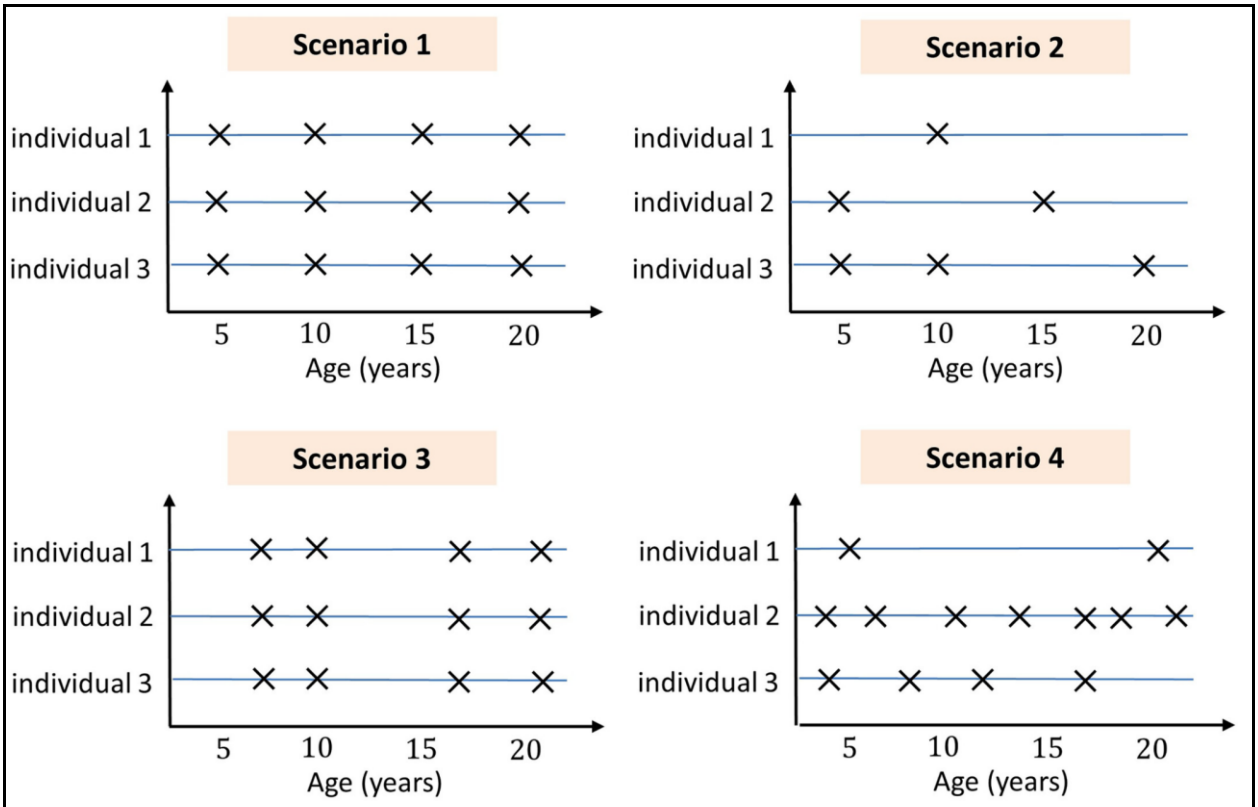


Figure 11. Examples of commonly seen longitudinal data structures[53]

2.3.1.3. Statistical mechanisms of missingness

Reasons for data missingness may or may not be recorded (observed). Missing data can be described by 1 of 3 mechanisms: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). When data is MCAR, the probability of it being missing is independent of any observed or unobserved data. Thus, while removing participants with MCAR data from analysis can reduce cohort size, it does not introduce bias. The probability of data MAR is dependent on observed data that is unrelated to the missing value itself. Removal of participants with this type of missingness introduces bias, but it can be dealt with in subsequent analysis since the relation is to an observed characteristic. Lastly, data that is MNAR has a missingness probability that is directly related to unobserved data. Therefore, MNAR data are

very difficult to characterize or even detect.[36,54] Mechanisms and patterns of data missingness have been discussed in more depth elsewhere.[55–57] Examples of structured missingness patterns in the context of health data are shown in Figure 12.

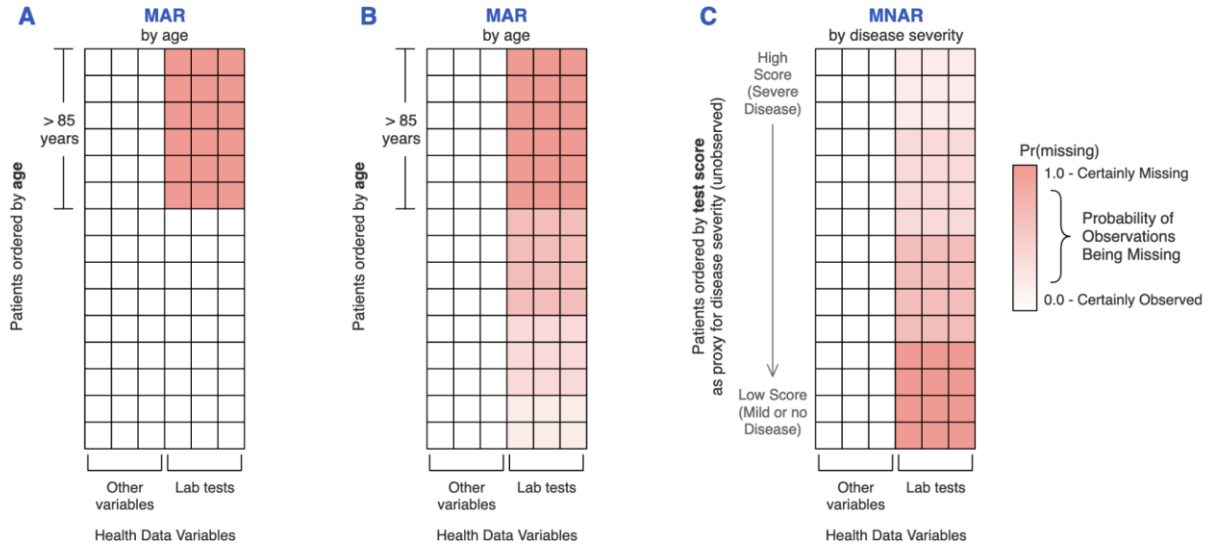


Figure 12. Examples of structured missingness[55]

2.3.2. Risk of bias

Real-world data sources can suffer from a number of biases, including confounding, selection bias, external validity bias, measurement (information) bias, and time-related bias, each of which has several subcategories.[42,58]

Confounding results from the choice of intervention and the outcome(s) having a common cause. For example, a worse disease status could be associated with more aggressive treatment but also poorer outcomes. In the MVTS context, confounding could manifest as sicker patients coming into the clinic more often and therefore having more repeat measurements available for analysis.

Selection bias occurs when the people under study are not representative of the target population. For multivariate trajectory analysis, including clustering, repeat measurements are required to form a time series. There may be nontrivial differences in the cohort of patients that has a sufficient number of repeat measurements compared to the unselected patients who do not. Selection bias can be caused by loss to follow up, if a subset of participants chooses to withdraw from a longitudinal observational cohort, for example. It can also be caused by excluding those participants with missing data from analysis. Selection bias can introduce, or further exacerbate, external validity bias, wherein the findings from a study are not as applicable as intended to (1) the population from which the sample was drawn and/or (2) another target population.[42]

Measurement (information) bias can result from missing or erroneous data collected on interventions, exposures, outcomes, covariates, and other patient features or characteristics. Such limitations may be a product of improper data collection processes, differing study protocols, or changing care patterns. These issues could be more or less consequential depending on their magnitude in a dataset, their randomness or systematicness, and their distribution across groups of interest. If present, this kind of bias could impact the interpretation of differences between identified clusters with regard to any variables that were improperly recorded.[42]

Lastly, time-related biases, which can be considered a subset of the other bias categories, occur when follow-up time and/or exposure status are insufficiently accounted for in the study design or analysis. A manifestation of this kind of bias is when treatments are given at different disease stages, inherently introducing disease duration- and progression-related biases, but comparisons do not account for that difference.[58]

2.3.3. High dimensionality

Not all available variables in a dataset may be related to underlying heterogeneity. Inclusion of uninformative variables can make it more difficult for a clustering method to function effectively. Pre-selection of variables that have high data quality and are most likely to capture heterogeneity between patients is beneficial.[34] Approaches to performing this selection are discussed in Section 4.4.2. It is also important to avoid over-representation of variables that measure the same construct, as this can lead to distances between clusters being influenced more heavily by one construct than another. However, some clustering approaches, like deep learning-based ones, are less susceptible to these issues.[36,59]

2.4. Deep representation learning for time series clustering

Time series clustering aims to partition a set of time series into a group of clusters by maximizing (1) the similarities between time series within the same cluster and (2) the dissimilarities between time series of different clusters (Figure 13).[38] Numerous methods have been devised to accomplish this task, and various taxonomies have been proposed to categorize them. Figure 14 shows one such taxonomy, devised by Paparrizos et al. Table 3 describes major clustering challenges and how different models may be suitable for overcoming them. These methods have been comprehensively described elsewhere.[35,53,60–66] Our work is solely concerned with deep representation learning-based methods (Figure 15). As data becomes increasingly complex, shallow (traditional) clustering methods struggle to handle high dimensionality. Deep representations can capture non-linear, additive, and multiplicative effects in lower dimensions, making them a natural extension to be used in clustering algorithms.[59,67]

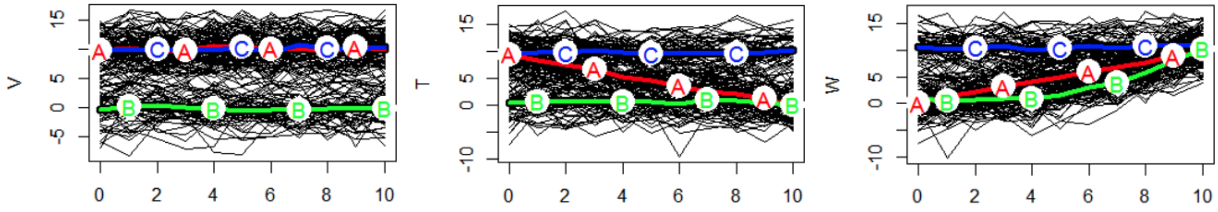


Figure 13. Identification of 3 clusters (A, B, C) across 3 longitudinal variables (V, T, W)[68]

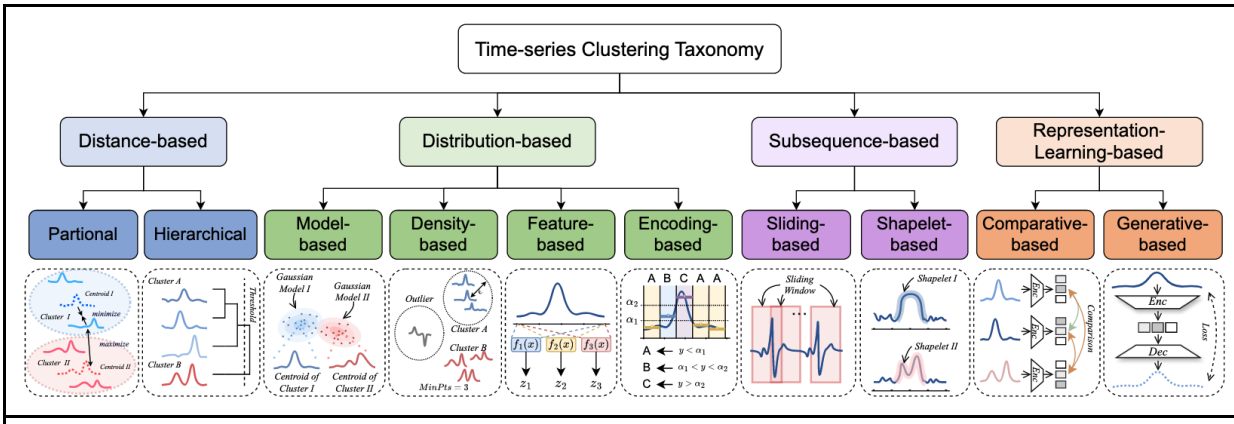


Figure 14. Taxonomy of time-series clustering algorithms proposed by Paparrizos et al.[27]

	Center-based clustering	Hierarchical clustering	Density-based clustering	Model-based clustering	Extensions
High dimensional, noise, and sparse data	Problematic	Problematic	Problematic	Problematic	Requires variable selection and dimensionality reduction or subspace clustering methods
Skewed distribution	Problematic	Problematic	Robust for models allowing density variation	Robust when distributions are correctly estimated	Data normalization is generally needed.
Outliers	Problematic	Problematic	Robust	Can be problematic	Fuzzy clustering is more robust to outliers. Outlier/anomaly detection models can be used prior to clustering.
Overlapping boundaries	Problematic	Problematic	Problematic	Robust	Fuzzy clustering can be used when there are overlapping cluster boundaries.
Arbitrary cluster shapes	Problematic	Likely to be problematic	Robust	Problematic	There are many extension models available such as kernel-based clustering and non-linear feature extraction models.
Rare events	Likely to be problematic	Likely to be problematic	Likely to be problematic	Likely to be problematic	Consider anomaly detection algorithms when aiming to identify very small clusters.
Mixed data	Potentially problematic	Potentially problematic	Potentially problematic	Potentially problematic	Distance measures for mixed data can be used, however can be sensitive to outliers and not capturing important features. Dimensionality reduction is needed prior to clustering.
Missing data	Problematic	Problematic	Problematic	Likely to be problematic	Multiple imputation models or clustering algorithms that specifically model data missingness are needed for data MAR.

Table 3. Challenges in clustering tasks and robustness for different models[59]

2.4.1. Deep representation learning

Deep neural networks (DNNs) can learn complex representations of data, making them suitable for capturing the intricacies of multivariate time series (high dimensionality, sparsity, non-linearity).[36,59,69] The goal of deep representation-based clustering methods for MVTS data is to (a) learn a lower-dimensional representation of the input MVTS that retains the information necessary to

compare it to others and (b) perform clustering on this new representation of the original data (Figure 16). Advances in deep learning have enabled these unsupervised learning approaches to achieve high performance.[27] In this section, we lay out common representation learning architectures, including deep neural network (DNN) layer types and loss functions, imputation strategies, and specific methods adapted for time series data.

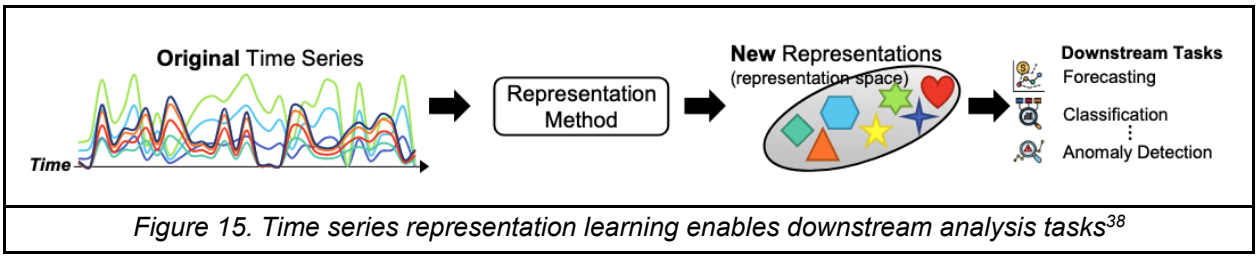


Figure 15. Time series representation learning enables downstream analysis tasks[38]

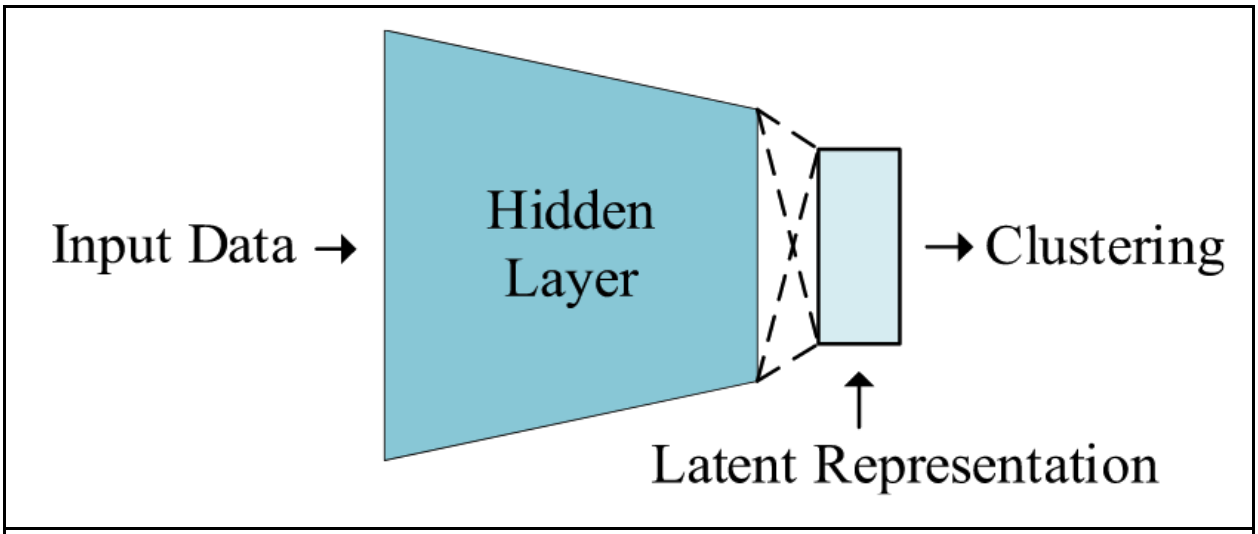


Figure 16. General framework of DNN-based deep clustering methods[70]

2.4.1.1. Encoder architectures

Deep neural networks (DNNs) are typically composed of an input layer, one or more hidden layers, and an output layer (Figure 17).[70] Each layer is a non-linear function that can be one of the following: Fully Connected Neural Network (FCNN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Attention-based Neural Network (Transformer), Graph Neural Network (GNN).[27] We focus our discussion on FCNN and RNN, as they are the most commonly used in our methods of interest.

FCNN, also known as Multilayer Perceptron (MLP), is the simplest DNN layer architecture. As the name implies, every neuron from one layer is connected to every neuron from the previous layer. FCNN have been applied widely across domains but struggle with capturing specific interrelations within a specific dimension, like space or time. RNNs are most commonly used for sequential data analysis due to their internal memory mechanism which allows them to take into account the temporal dimension of the input.[27,28] Figure 17 shows the difference between FCNN and RNN. The RNN's characteristic recursive behavior, denoted by the circular arrow, is "unrolled" in Figure 18. However, the standard RNN struggles with long-term dependencies (remembering information for long periods of time). The Long Short Term Memory (LSTM) variant of the standard RNN was proposed to achieve better long-term recall. LSTM was further iterated on to produce other variants, including peephole LSTM and Gated Recurrent Unit (GRU).[71] These types of recurrent cells are shown in Figure 19. Notably, no single variant consistently outperforms the others.[71,72]

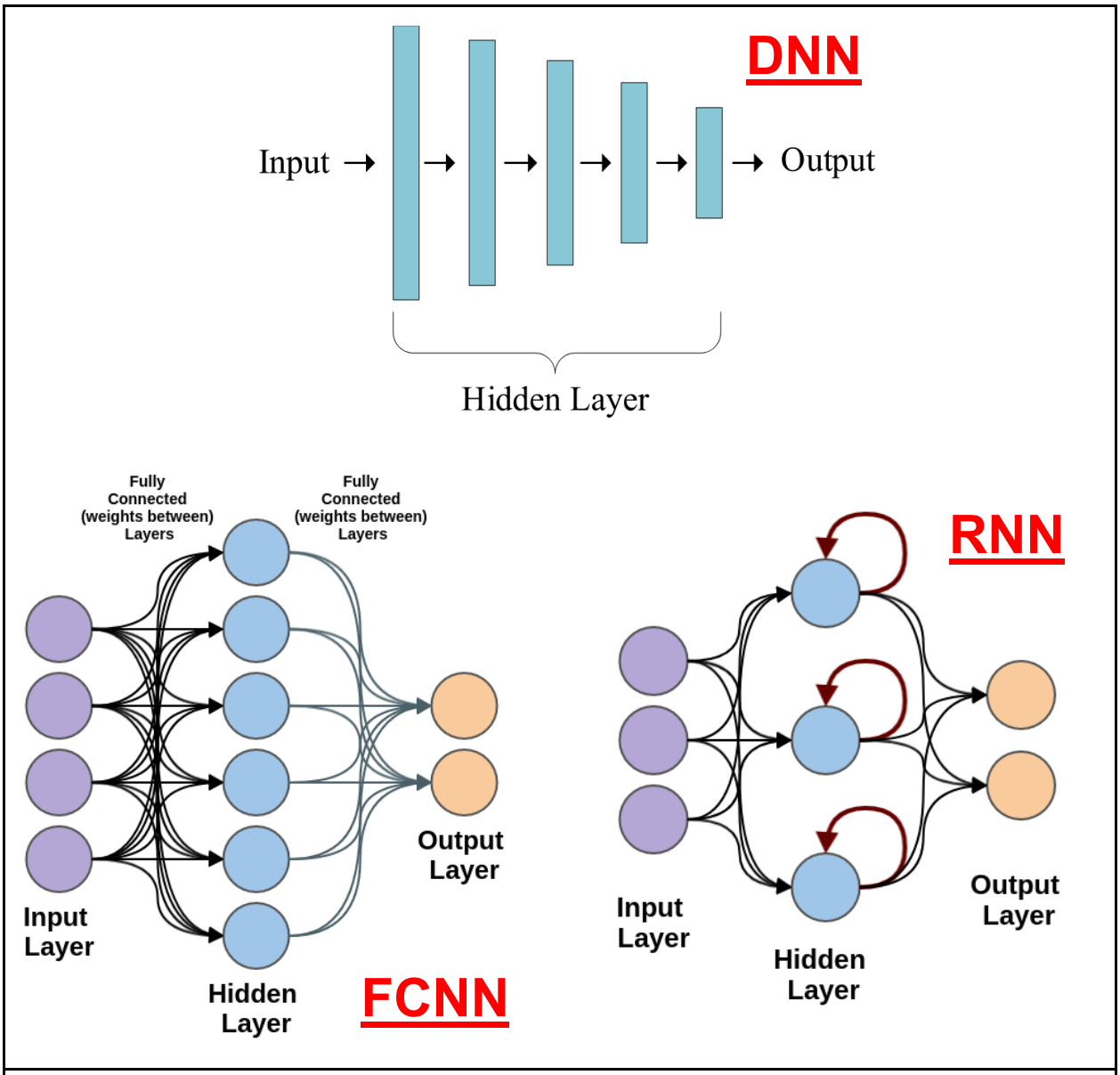


Figure 17. General DNN vs. FCNN vs. RNN[70,73]

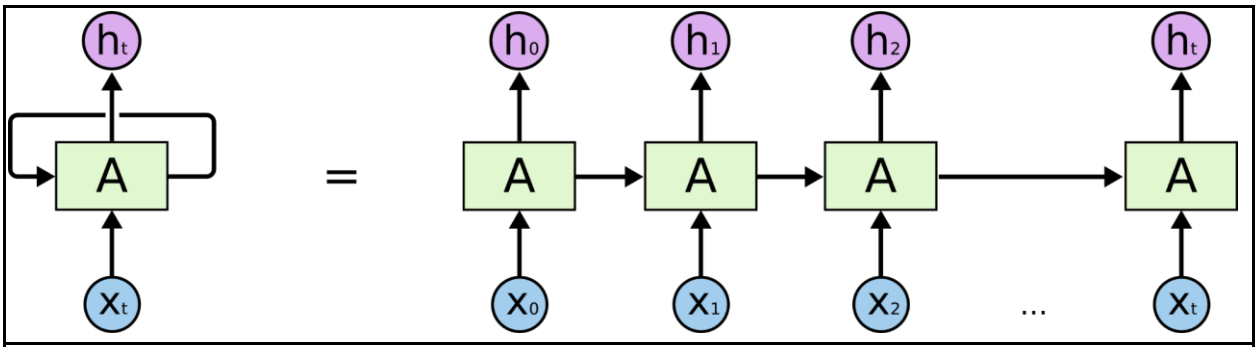


Figure 18. Unrolled RNN unit[71]

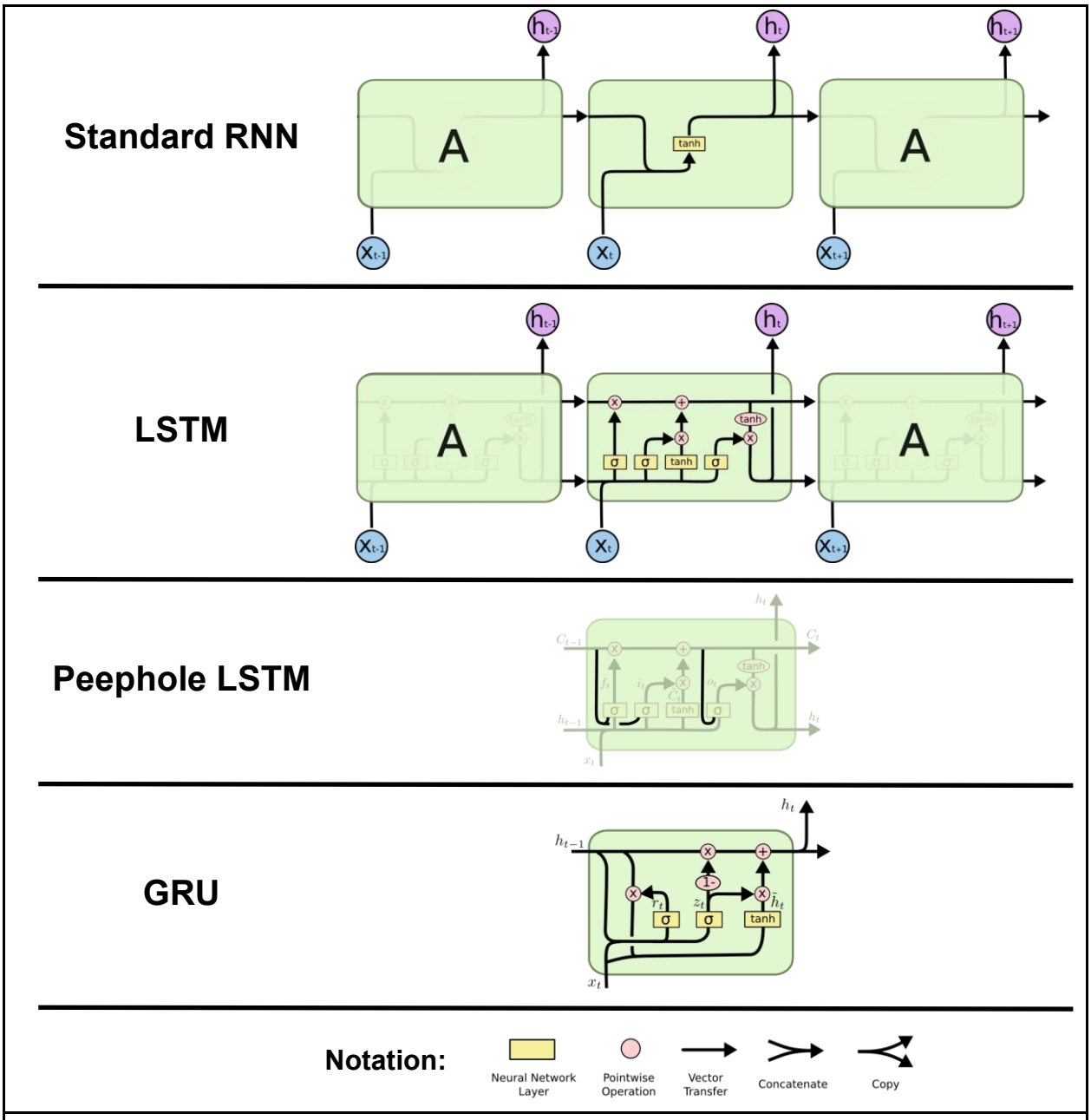


Figure 19. Types of recurrent units[71]

2.4.1.2. Pretext losses

To enable a DNN to learn a lower-dimensional, yet faithful, representation of the input MVTS, a number of pretext losses have been proposed. This is one of the objective functions that the DNN is optimized on. The most common is the reconstruction loss associated with the autoencoder (AE) architecture (Figure 20). The autoencoder learns meaningful features in the latent

representation by minimizing the error between the original input and the reconstructed output.[27] The variational autoencoder (VAE) is a widely-used AE variant that improves the capability of generalizing to new data (Figure 21). The VAE encoder maps inputs to a probability distribution in the latent space (as opposed to individual points, like the AE encoder does). The total VAE loss function, known as evidence lower bound (ELBO), is achieved by adding a Kullback-Leibler (KL) divergence constraint to the reconstruction loss.[70] Other pretext losses include triplet loss, InfoNCE loss, and cross-entropy loss.[28] Figure 22 shows an example of an autoencoder architecture that uses RNN layers to process multivariate time series for anomaly detection.

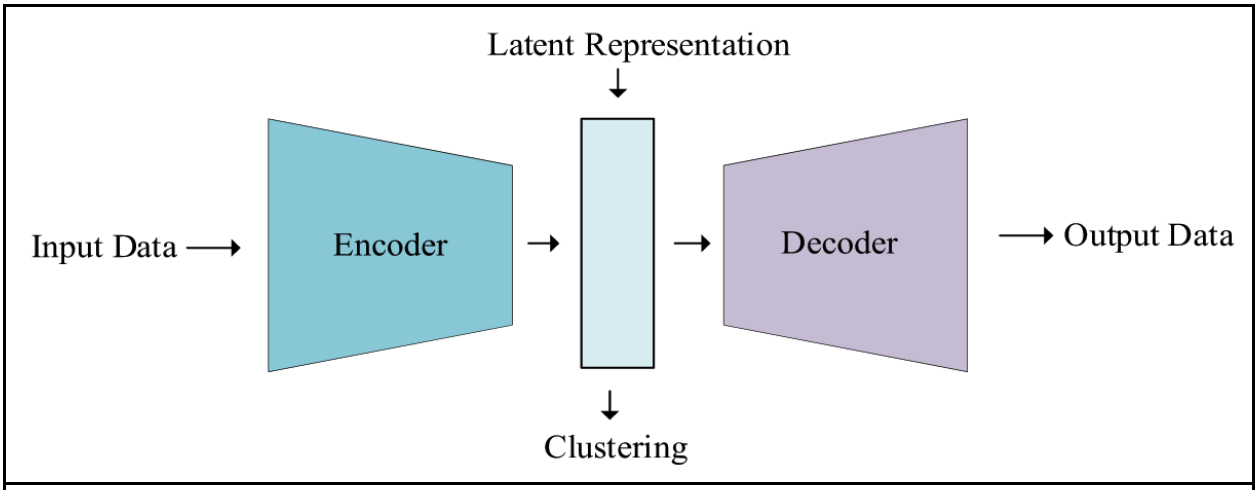


Figure 20. General framework of AE-based deep clustering methods[70]

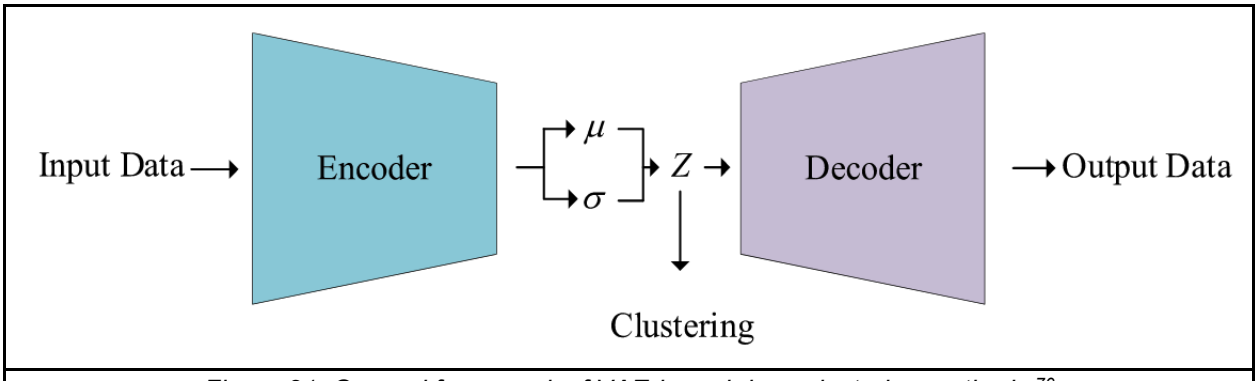


Figure 21. General framework of VAE-based deep clustering methods.[70]

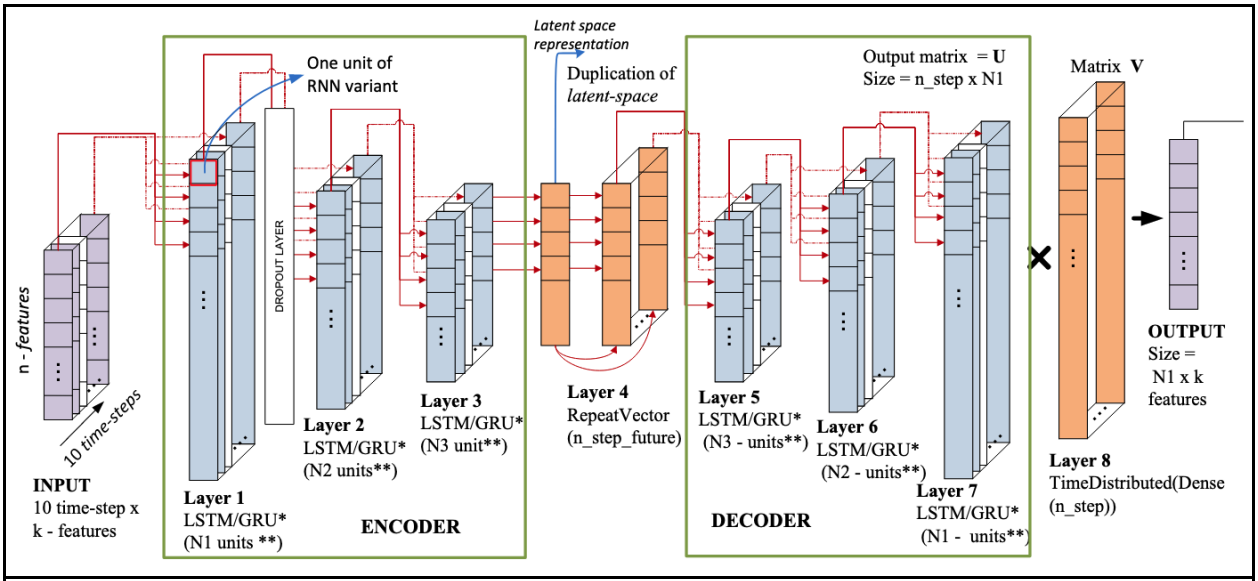


Figure 22. Architecture design of RNN-Autoencoder[72]

2.4.1.3. Clustering losses

A pretext loss alone may not be sufficient to obtain a latent space that is suitable for clustering. To encourage learning of a more separable latent space, a number of clustering losses have been proposed as a complementary objective to be optimized alongside the pretext loss.[28] Commonly used clustering losses include DEC, IDEC, DEPICT, SDCN, VaDE, DTCR, and ClusterGAN (Table 4).[27,28]

Authors	Year	Full name	Abbr.	Based on
Xie, Girshick, Farhadi	2016	Deep Embedded Clustering	DEC[74]	
Guo, Gao, Liu, Yin	2017	Improved Deep Embedded Clustering	IDEC[75]	DEC
Jiang, Zheng, Tan, Tang, Zhou	2017	Variational Deep Embedding	VaDE[76]	
Dizaji, Herandi, Deng, Cai, Huang	2017	DEeP Embedded RegularIzed ClusTering	DEPICT[77]	DEC
Mukherjee, Asnani, Lin, Kannan	2019	Cluster Generative Adversarial Network	ClusterGAN[78]	
Ma, Zheng, Li, Cottrell	2019	Deep Temporal Clustering Representation	DTCR[79]	IDEC
Bo, Wang, Shi, Zhu, Lu, Cui	2020	Structural Deep Clustering Network	SDCN[80]	IDEC

Table 4. Widely used clustering losses[27]

DEC is one of the most referred to and has been adapted and refined by many. It introduces an auxiliary target distribution against which a soft clustering assignment distribution is assessed

using Kullback-Leibler (KL) divergence (Figure 23). VaDE combines VAE and a Gaussian Mixture Model to learn as many distributions as expected clusters.[b] Samples are generated for each distribution and the network is trained to reconstruct them and fit the generated distribution with the original (Figure 24). DTCR incorporates an additional classification task to train the encoder to discriminate between real and fake input. It also introduces a k-Means clustering loss (Figure 25).[27,28]

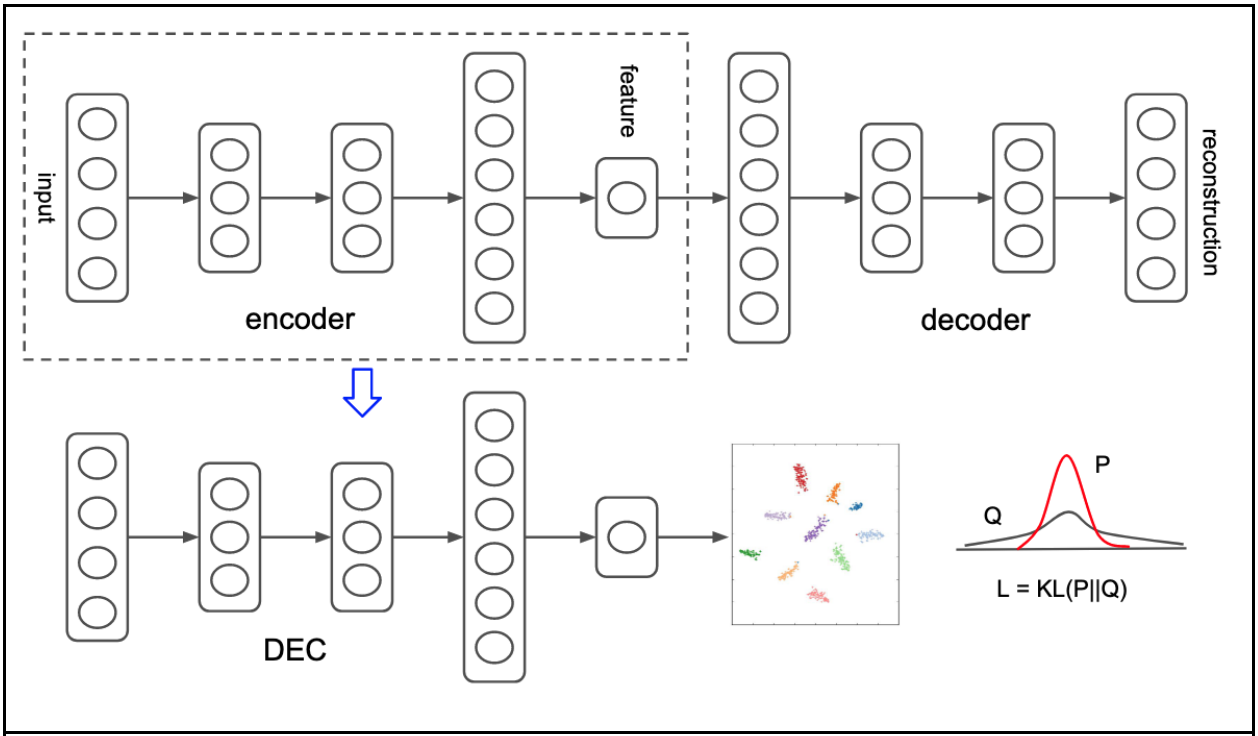


Figure 23. DEC structure[74]

[b] In Section 2.4.3.1, we discuss Variational Deep Embedding with Recurrence (VaDER), which extends VaDE to the time series setting. VaDER is one of the methods studied deeply in Aim 3.

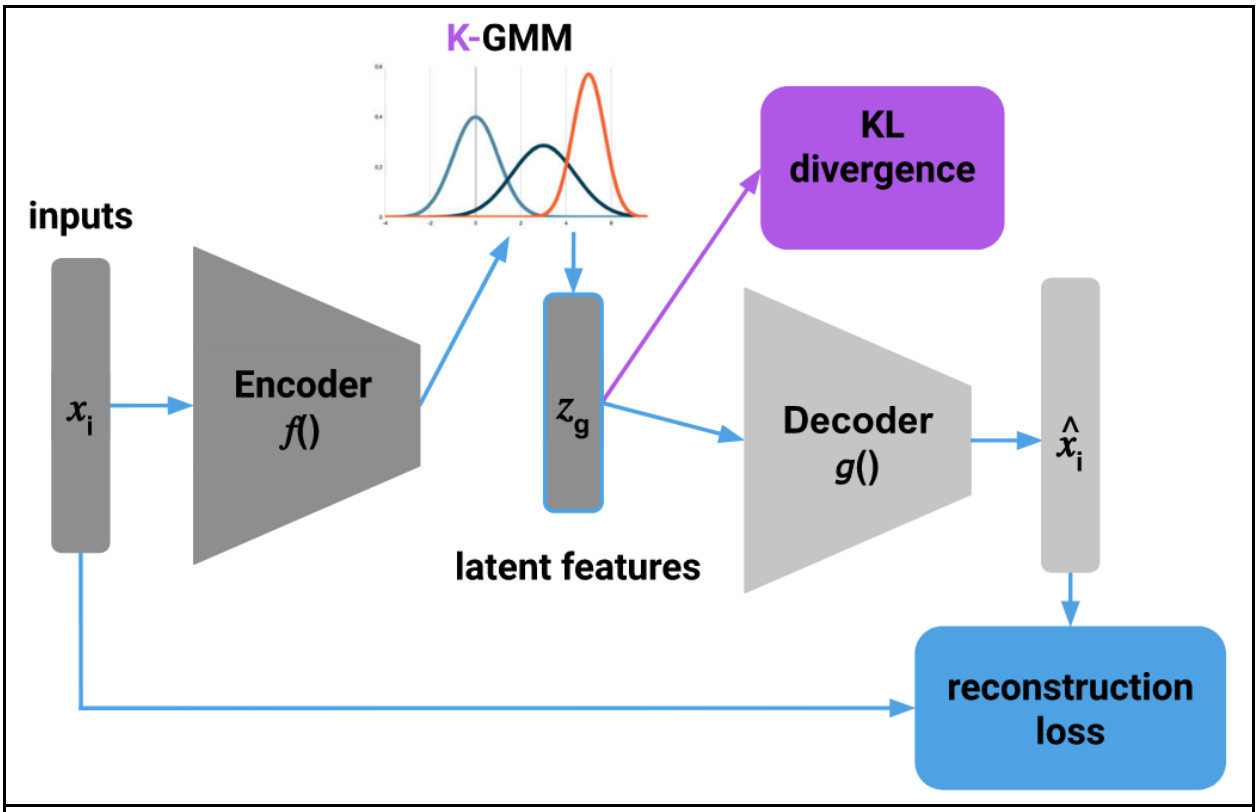


Figure 24. VaDE method with K clusters[28]

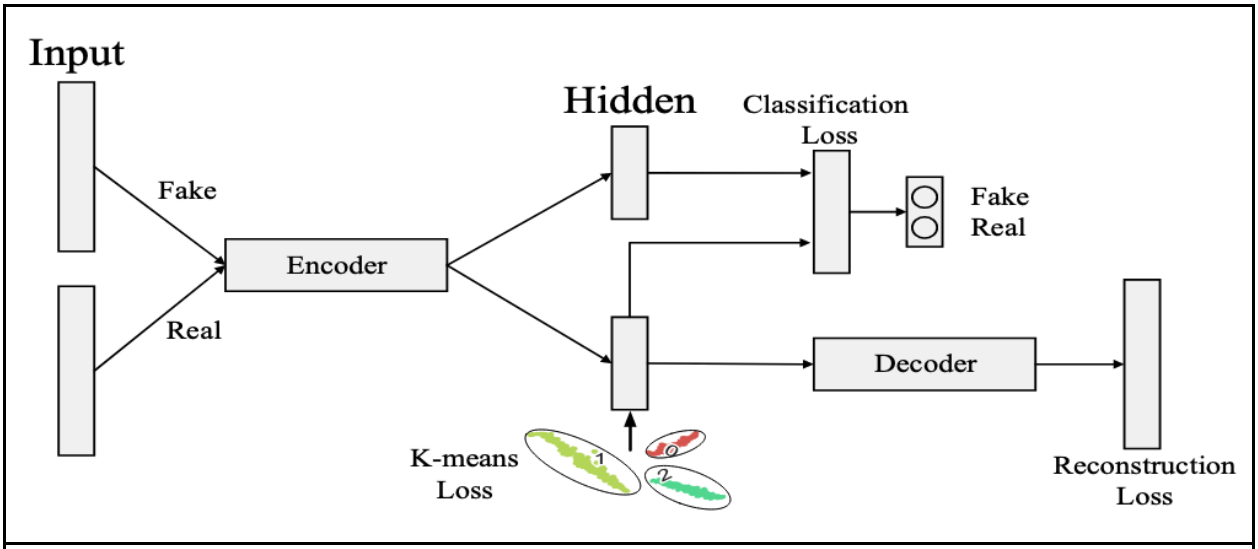


Figure 25. DTCR architecture[79]

2.4.1.4. Summary of methods

Deep time series clustering methods have been catalogued and benchmarked by several others.[26–29] Figure 26 shows a taxonomy of time series clustering methods, include deep learning-

based ones, published by Paparrizos et al. in 2025. As discussed in Section 5.3.1, benchmarks to date have focused on univariate and/or non-missing MVTS datasets.

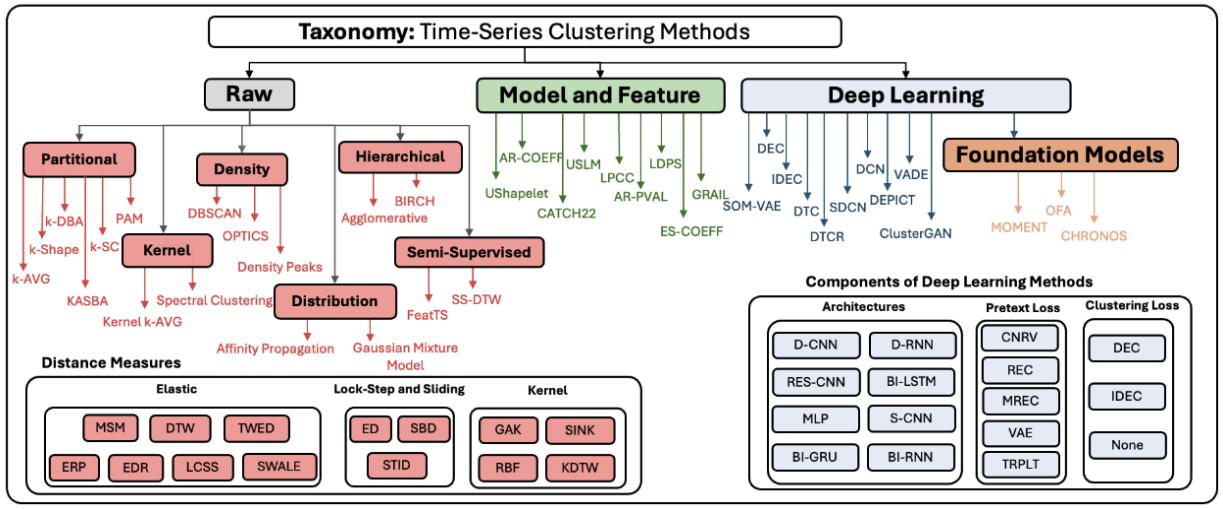


Figure 26. Taxonomy of time series clustering methods[29]

2.4.2. MVTS imputation

The methods discussed thus far cannot natively handle incomplete time series, so imputation is typically performed prior to clustering. The simplest approach is to replace missing values with zero, mean, or median values.[33] More sophisticated deep learning-based imputation methods have been discussed in depth elsewhere.[52,56,81–83] Figure 27 shows a taxonomy of such methods with one representative model per category.

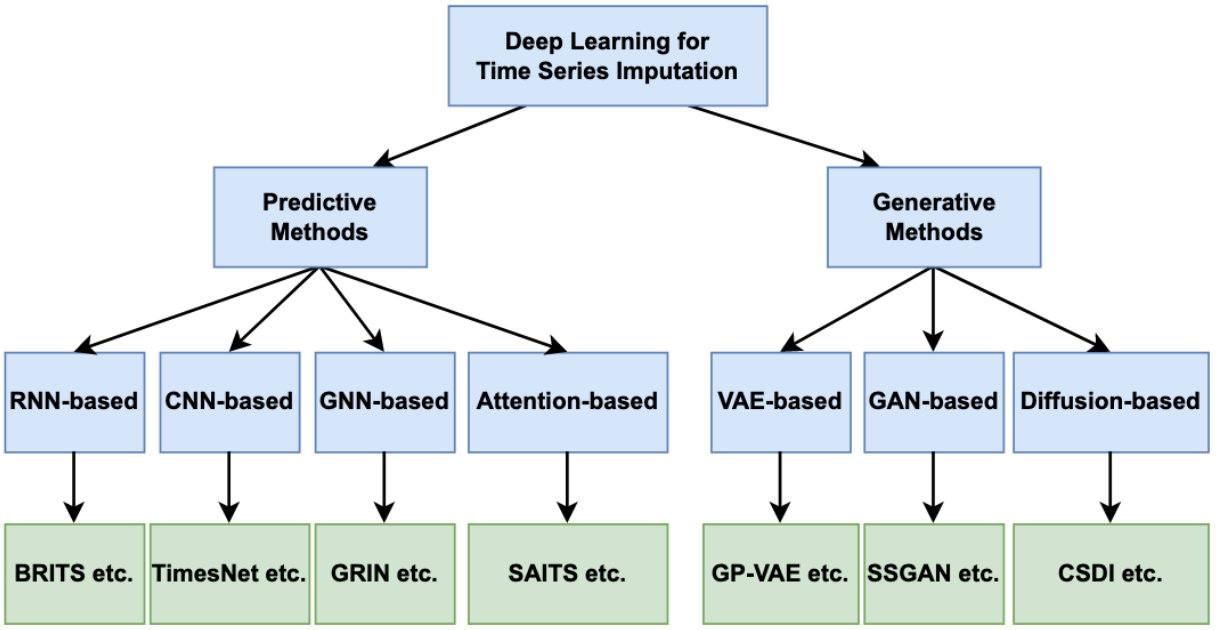


Figure 27. Taxonomy of deep learning-based MVTS imputation methods[81]

2.4.3. One-stage MVTS clustering methods

Missing value handling in time series clustering is divided into two categories: two-stage approaches and one-stage (end-to-end) approaches. Two-stage methods impute first, cluster second. However, this can lead to suboptimal results because the missingness patterns themselves are not learned from by the model.[52] To rectify this, one-stage methods jointly optimize imputation and clustering.[33,36] Methods that integrate MVTS deep representation learning, clustering, and imputation into one end-to-end framework include VaDER and CRLI (Figure 28).

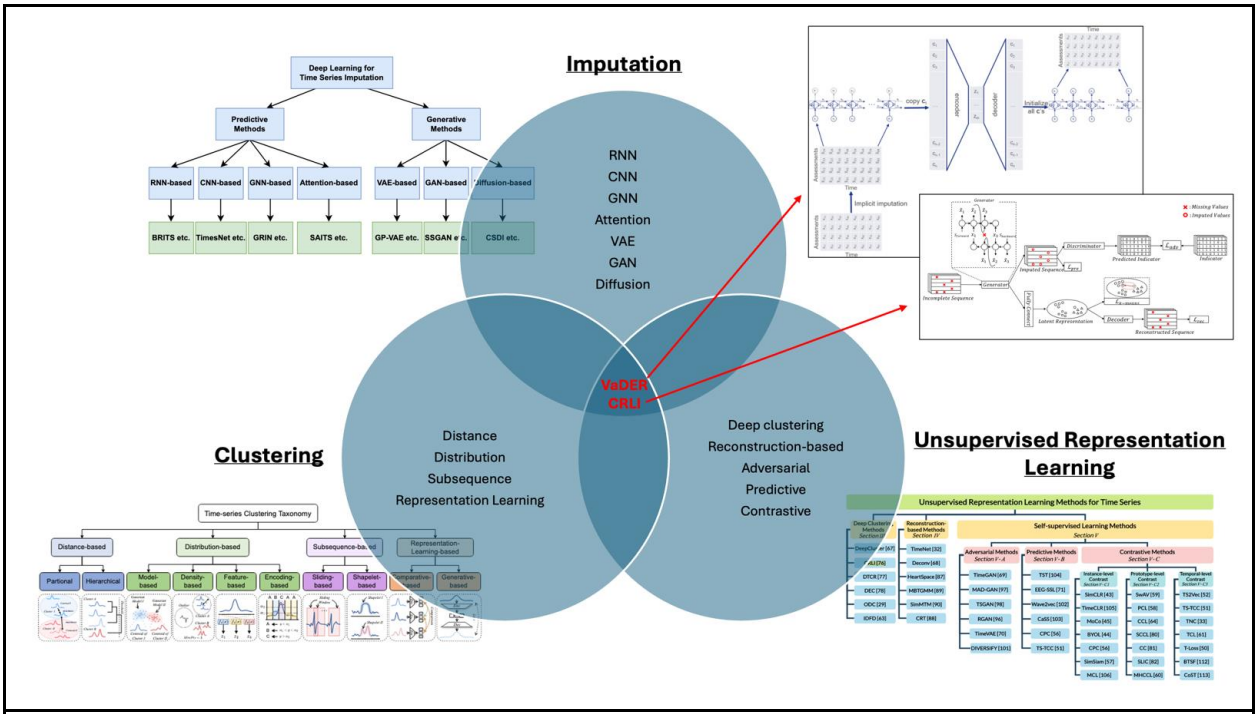


Figure 28. One-stage deep representation learning for MVTS clustering: VaDER and CRLI[27,33,54,69,81]

2.4.3.1. Variational Deep Embedding with Recurrence (VaDER)

VaDER is an autoencoder-based one-stage MVTS clustering method published in 2019 by de Jong et al.[54] It uses the VaDE (Figure 24) latent loss to simultaneously learn latent representations and cluster assignments of its input samples. It also incorporates peephole LSTM (Figure 19) networks into the autoencoder architecture to model the auto- and cross-correlations in the MVTS data (Figure 29). VaDER includes an implicit imputation scheme which is performed within model training. This is achieved by defining a weighted reconstruction loss wherein imputed values are weighted to 0, non-imputed values to 1. Missing values are initially imputed with arbitrary values then treated as trainable parameters which are iteratively updated by stochastic gradient descent. VaDER benchmarking experiments are described in Section 5.3.2. To date, VaDER has been used to identity progression profiles of demented Alzheimer's disease patients and Parkinson's disease patients.[30,31]

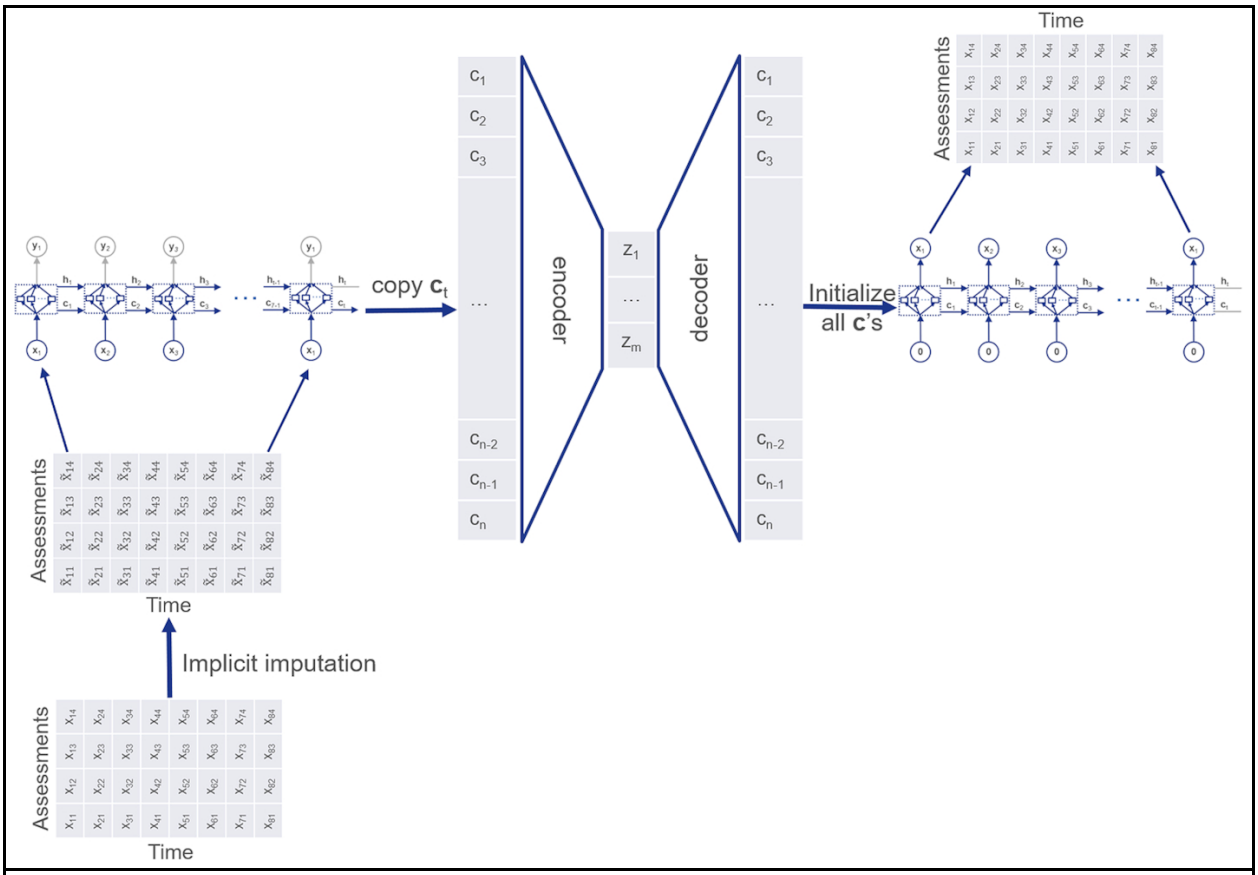


Figure 29. VaDER architecture[54]

2.4.3.2. Clustering Representation Learning on Incomplete time-series data (CRLI)

CRLI, published in 2021 by Ma et al., is composed of two parallel branches which share a bidirectional RNN acting as a generator for the upper branch and encoder for the lower branch (Figure 30). In the upper branch, an adversarial strategy is employed to reduce error propagation from imputation to clustering. The lower branch integrates a soft K-means objective (as in Figure 25) to produce cluster-amenable representations. This joint training strategy makes the imputed values more acceptable for clustering. As annotated in blue text in Figure 30, the generator/encoder is a bidirectional multi-layer RNN, the decoder is a single-layer RNN, and the discriminator is a 5-layer RNN. CRLI uses Gated Recurrent Units (GRU) in the RNN, as opposed to LSTM (Figure 19).[33] To date, CRLI has been used to identity distinct groups of clinical trajectories of septic patients with acute respiratory failure in the ICU.[32]

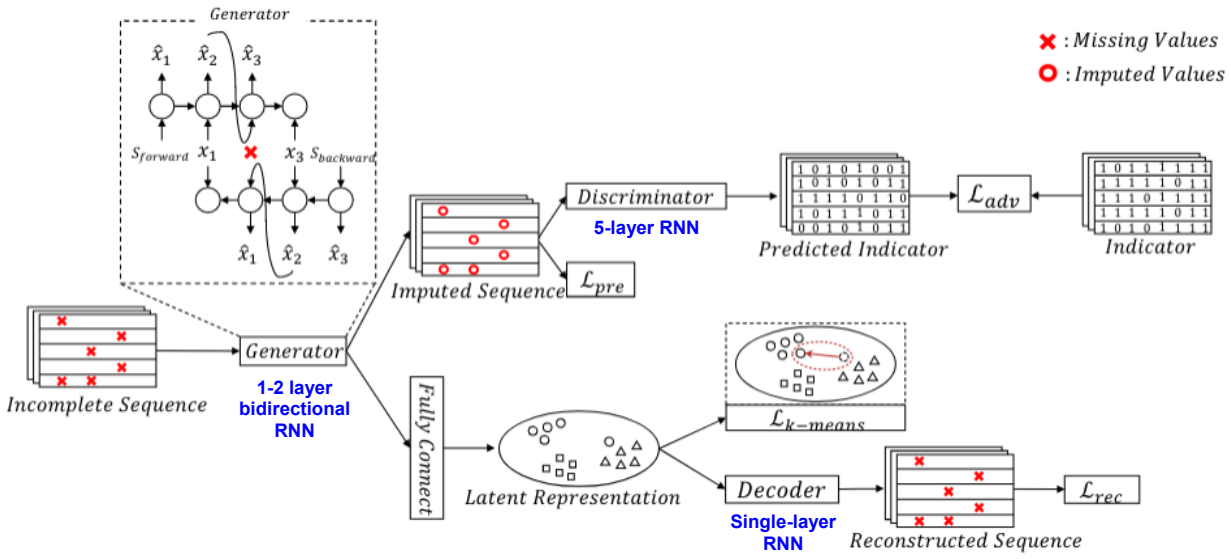


Figure 30. CRLI architecture[33]

2.4.3.3. Comparative analysis

As discussed in Section 5.3.2, Ma et al. compared CRLI performance to VaDER's on 8 real-world datasets (Table 19). They also tested several two-stage methods, first imputing by using one of three imputation methods (ZERO, GAIN, BRITS), then applying one of five clustering methods (k-Shape, DEC, IDEC, DTC, DTCR).[c] These imputation and clustering methods are referenced in Figure 27 and Table 4, respectively. Though CRLI outperformed VaDER and the two-stage methods on most datasets (6/8) and metrics tested, VaDER was competitive on the two datasets which were multivariate and sparse. Ma et al. did acknowledge this but attributed the performance to the relatively larger training sizes of those two datasets compared to the others, stating that VaDER requires more data owing to its generative clustering framework. They also critiqued VaDER for not having constraints on missing values on the encoder side, likely leading to errors being introduced into the clustering process.[33]

[c] ZERO = zero imputation; GAIN = Generative Adversarial Imputation Nets; BRITS = Bidirectional Recurrent Imputation for Time Series; DEC = Deep Embedded Clustering; IDEC = Improved Deep Embedded Clustering; DTC = Deep Temporal Clustering; DTCR = Deep Temporal Clustering Representation

2.4.4. Clustering evaluation

For any cluster analysis, MVTS or otherwise, careful evaluation of the clustering result is paramount before drawing conclusions. Generally, this is made difficult by the lack of ground truth labels in an unsupervised clustering scenario. Approaches to evaluation include (1) internal clustering validation indices (CVI), when ground truth is not available, (2) external CVI, when ground truth is available, (3) manual evaluation by a human domain expert, and/or (4) indirect evaluation of the utility of the clustering result in its application area.[36]

2.4.4.1. Internal clustering validation indices

Internal CVI quantify the the goodness of a clustering structure without referring to external information (like ground truth labels). They can be used to choose (1) between clustering algorithms and (2) the optimal cluster number for a given dataset. Internal CVI are often based on two criteria: compactness and separation. The former measures how closely related objects in the same cluster are to one another. The latter measures how distinct a cluster is from the others. 11 widely used internal CVI are shown in Table 5.[84] Multiple studies have shown that no single index dominates in every situation.[85,86] Therefore, we chose 4 indices (highlighted in Table 5) that are well represented in the time series literature.[68,87]

	Measure	Notation	Definition	Optimal value
1	Root-mean-square std dev	$RMSSTD$	$\{\sum_i \sum_{x\in C_i} \parallel x-c_i \parallel^2/[P\sum_i(n_i-1)]\}^{\frac{1}{2}}$	Elbow
2	R-squared	RS	$(\sum_{x\in D} \parallel x-c \parallel^2 - \sum_i \sum_{x\in C_i} \parallel x-c_i \parallel^2)/\sum_{x\in D} \parallel x-c \parallel^2$	Elbow
3	Modified Hubert Γ statistic	Γ	$\frac{2}{n(n-1)} \sum_{x\in D}\sum_{y\in D} d(x,y)d_{x\in C_i, y\in C_j}(c_i,c_j)$	Elbow
4	Calinski-Harabasz index	CH	$\frac{\sum_i n_i d^2(c_i,c)/(NC-1)}{\sum_i \sum_{x\in C_i} d^2(x,c_i)/(n-NC)}$	Max
5	I index	I	$(\frac{1}{NC}\cdot \frac{\sum_{x\in D} d(x,c)}{\sum_i \sum_{x\in C_i} d(x,c_i)} \cdot \max_{i,j} d(c_i,c_j))^p$	Max
6	Dunn's indices	D	$\min_i\{\min_j(\frac{\min_{x\in C_i, y\in C_j} d(x,y)}{\max_k\{\max_{x,y\in C_k} d(x,y)\}})\}$	Max
7	Silhouette index	S	$\frac{1}{NC}\sum_i\{\frac{1}{n_i}\sum_{x\in C_i}\frac{b(x)-a(x)}{\max[b(x),a(x)]}\}$ $a(x)=\frac{1}{n_i-1}\sum_{y\in C_i, y\neq x} d(x,y), b(x)=\min_{j,j\neq i}[\frac{1}{n_j}\sum_{y\in C_j} d(x,y)]$	Max
8	Davies-Bouldin index	DB	$\frac{1}{NC}\sum_i \max_{j,j\neq i}\{[\frac{1}{n_i}\sum_{x\in C_i} d(x,c_i)+\frac{1}{n_j}\sum_{x\in C_j} d(x,c_j)]/d(c_i,c_j)\}$	Min
9	Xie-Beni index	XB	$[\sum_i \sum_{x\in C_i} d^2(x,c_i)]/[n\cdot min_{i,j\neq i} d^2(c_i,c_j)]$	Min
10	SD validity index	SD	$Dis(NC_{max})Scat(NC)+Dis(NC)$ $Scat(NC)=\frac{1}{NC}\sum_i \parallel \sigma(C_i) \parallel / \parallel \sigma(D) \parallel, Dis(NC)=\frac{max_{i,j} d(c_i,c_j)}{min_{i,j} d(c_i,c_j)}\sum_i(\sum_j d(c_i,c_j))^{-1}$	Min
11	S_Dbw validity index	S_Dbw	$Scat(NC)+Dens\_bw(NC)$ $Dens\_bw(NC)=\frac{1}{NC(NC-1)}\sum_i[\sum_{j,j\neq i}\frac{\sum_{x\in C_i\cup C_j} f(x,u_{ij})}{max\{\sum_{x\in C_i} f(x,c_i), \sum_{x\in C_j} f(x,c_j)\}}]$	Min

D: data set; n: number of objects in D; c: center of D; P: attributes number of D; NC: number of clusters; C_i: the i–th cluster; n_i: number of objects in C_i; c_i: center of C_i; $\sigma(C_i)$: variance vector of C_i; $d(x,y)$: distance between x and y; $\parallel X_i \parallel = (X_i^T \cdot X_i)^{\frac{1}{2}}$

Table 5. Internal Clustering Validation Measures[84]

2.4.4.2. External clustering validation indices

External CVI compare clustering results with a known underlying class labeling. They are used in evaluation frameworks to compare method performance and estimate stability of cluster assignments over resampled data. External CVI are typically classified into three categories according to their similarity assessment technique: pair-counting, information theoretic, and set-matching.[88] Several analyses have cemented Adjusted Rand Index (ARI) as the standard for external evaluation due to its independence from the number of clusters and adjustment for chance.[88–90] In Aim 3, we compute ARI alongside three other commonly reported metrics (purity, RI, NMI), equations for which are taken from Paparrizos et al. and shown in Table 6.[28,54,89] Other metrics not discussed here include F-measure, Entropy, Adjusted Mutual Information (AMI), Fowlkes Mallows index, Homogeneity, Completeness, and Cluster Accuracy.[27,89]

Measure	**Notation**	**Definition**	**Range**
Purity	Purity	$\frac{1}{N}\sum_{i}\max_{j}\left\|C_i \cap T_j\right\|$ where N represents the total number of time-series data points, C_i represents the cluster i, and T_j represents for the ground truth class label j	[0,1]
Rand Index	RI	Given two clusterings generated on the same data X, the predicted clustering C and the ground truth T, RI is computed as follows: $\frac{TP+TN}{TP+TN+FP+FN}$ where True Positives (TP) represents the frequency with which data points in a point-pair are grouped within the same cluster in both C and T, True Negatives (TN) represents the frequency with which data points in a point-pair are assigned to different clusters in both C and T, False Positives (FP) represents the count of occurrences where data points in a point-pair are clustered to the same cluster in C while T separated them into different clusters, and False Negatives (FN) represents the count of occurrences where data points in a point-pair are separated into the different clusters in C while T grouped them into the same cluster	[0,1]
Adjusted Rand Index	ARI	$\frac{RI - Expected\ RI}{Maximum\ RI - Expected\ RI}$	[-1,1]

Normalized Mutual Information	NMI	Given two clusterings M and N on the same set of data points X, where M and N can have different numbers of clusters, NMI is computed as follows: $$NMI(M,N) = \frac{I(M;N)}{\sqrt{H(M)\cdot H(N)}},$$ $$I(M;N) = H(N) - H(N\|M),$$ $$H(M) = -\sum_{m\in M} P(m)\log_2 P(m),$$ $$H(N) = -\sum_{n\in N} P(n)\log_2 P(n),$$ $$H(N\|M) = -\sum_{m\in M, n\in N} P(m,n)\log_2 \frac{P(m,n)}{P(m)}$$ where $P(m, n)$ represents the joint probability of the value of M is m and the value of N is n while $P(m)$ and $P(n)$ represents the probability of M taking on the value m and the probability of N taking on the value n respectively	[0,1]

Table 6. Selected external clustering validation indices and their equations[27]

2.4.4.3. Visualization

Visualization can serve as a valuable qualitative tool for explaining and evaluating results. The latent space can be viewed to better understand aspects of cluster structure, like compactness and separation. Applying dimensionality reduction methods like t-SNE or UMAP, as others have done, can generate a 2D plot where clusters are more easily decipherable than in the multivariate longitudinal space. Such plots can serve as a complement to quantitative metrics to instill confidence that a sensible separation has been found by the model. [32,33,91] A mock 2D visualization of clustered MVTS is shown in Figure 31.

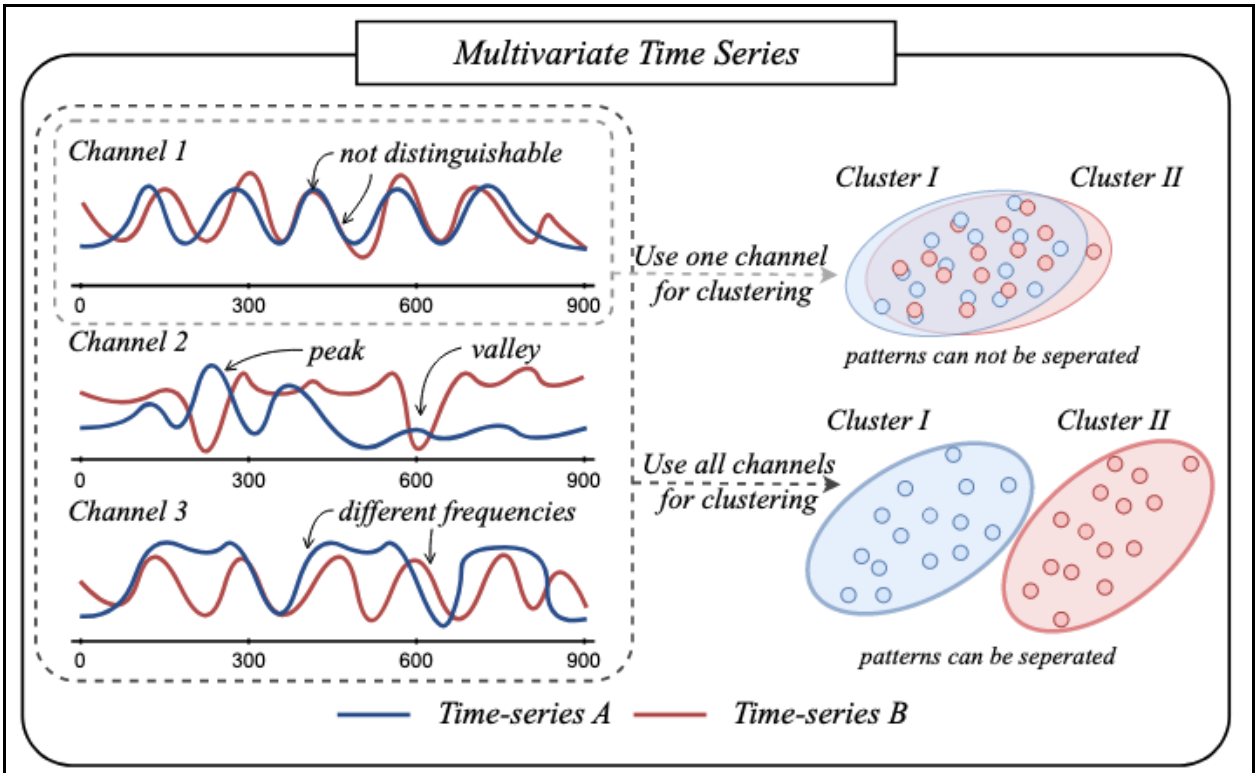


Figure 31. 2D representation of clustered MVTS data[27]

2.5. Prior work and papers related to specific Aims

Thus far, we have covered topics that are relevant to all Aims. Aims 1 and 2 are each structured around a specific longitudinal biomedical data source (retrospective EHR, prospective observational cohort) and trajectory subphenotyping task for CRLI to tackle (GLP-1 RA response, adolescent development). Aim 3 is concerned with the design of a synthetic MVTS dataset generation framework and evaluation of VaDER and CRLI using such generated datasets. Discussion of prior work and papers related to topics specific to each Aim can be found in their respective chapters. Section 3.2 and Section 3.3 discuss responder endpoints in clinical trials, the NIH *All of Us* dataset, treatment response modeling , GLP-1 RA effects, and trajectories of diabetes progression and GLP-1 RA trends derived from the longitudinal EHR. Section 4.2 and Section 4.3 discuss the multifaceted nature of adolescent development, the Adolescent Brain Cognitive Development (ABCD) Study, and self-injurious behavior and pubertal timing during

adolescence. Section 5.2 and Section 5.3 discuss considerations for machine learning benchmark datasets, properties of longitudinal biomedical datasets, VaDER and CRLI benchmarking performed to date, and approaches to synthetic time series generation.

2.6. Gaps in current knowledge and unanswered questions

Given our interest in identifying trajectory subtypes in diverse longitudinal biomedical datasets, which are generally sparse, high-dimensional, and temporally complex, we isolated VaDER and CRLI as being the most suitable methods available. We observed that CRLI had yet to be deployed in a novel biomedical context (outside of the 2 medical benchmark datasets it was evaluated on, which had ground truth labels), especially one characterized simultaneously by (i) short time series length (<15 timepoints), (ii) small cohort size (<700), and (iii) multiple variables (Table 19). We also observed that, while CRLI was more performant than VaDER on first glance, it had not been comprehensively evaluated against VaDER on a synthetic benchmark with a wide variety of time series and dataset characteristics. We devised our Aims to address these methodological gaps, and satisfy our own biomedical interests. Our overarching research question was thus: How and when are one-stage MVTS clustering methods (VaDER, CRLI) useful in biomedical research data exploration?

To that end, in Aim 1 (*Assessing the ability of CRLI to detect meaningful trajectories in a sparse, irregular, biased real-world dataset*, Chapter 3) we assessed the ability of CRLI to detect meaningful trajectories of GLP-1 RA response from short, sparse, irregularly sampled time series from the *All of Us* EHR. In Aim 2 (*Assessing the ability of CRLI to detect meaningful trajectories in a high-dimensional, multimodal, prospective dataset*, Chapter 4), we assessed the ability of CRLI to detect meaningful trajectories of adolescent development from short, multidomain, prospectively specified time series from the ABCD Study. And lastly, in Aim 3 (*Assessing the ability of CRLI and VaDER to detect trajectories in synthetic datasets under diverse data*

constraints, Chapter 5), we assessed head-to-head VaDER and CRLI performance on 20 synthetic MVTS datasets of our own creation. In the process, we formalized a framework that allows for rapid generation of novel time series models for 7 distinct variable styles, which can be combined to make novel datasets.

In the following chapter (Chapter 3: *Assessing the ability of CRLI to detect meaningful trajectories in a sparse, irregular, biased real-world dataset*), we begin our exploration by deploying CRLI in the temporal EHR to understand longitudinal patterns of GLP-1 RA treatment response across 5 routinely collected clinical measures.

3. Aim 1: Assessing the ability of CRLI to detect meaningful trajectories in a sparse, irregular, biased real-world dataset (NIH *All of Us* EHR)

3.1. Introduction

We began our experimental work with an assessment of CRLI in the context of longitudinal electronic health records (EHRs). This data source is unique in that it most faithfully captures a patient's experience of clinical care. The fullness of the medical history contained in the EHR, which can span heterogeneous types of data (labs, physical measurements, patient-reported outcomes, free text, images), provides a comprehensive impression of each patient. Though not originally intended for, or structured around, downstream research use, EHRs have been increasingly used to complement other biomedical data sources for hypothesis generation, pharmacovigilance, target trial emulation, treatment effects assessment, and more.[92–95] Along with claims and billing activities, disease registries, and patient-generated data, EHRs fall under the larger umbrella of real-world data (RWD), which is data collected during the routine delivery of health care.[96]

Temporal EHR data can be mined to perform longitudinal analyses that go beyond the individual-level to enable population-level knowledge discovery. However, this temporality comes with specific difficulties, like irregular time series length and variable measurement missingness within and across features and patients, which traditional machine learning approaches may be unequipped for.[46] In response, advanced deep learning techniques, and self-supervised representation learning techniques in particular, have emerged as a powerful and performant analytical solution.[97,98] Of these, the major contribution that one-stage, or end-to-end, methods, like CRLI and VaDER, make to the area of deep MVTS clustering is that they jointly optimize imputation and clustering. Imputation is the process of replacing missing data points with substituted values, and clustering is the process of grouping points in a dataset, or patients in a

cohort, based on degrees of similarity.[54] One-stage methods accept incomplete, or "partially observed", time series as inputs without requiring pre-imputation, in contrast to two-stage methods, which involve first imputing the incomplete time series dataset and then applying a clustering technique to the resultant complete dataset.[33,99,100] End-to-end MVTS clustering strategies that minimize error propagation from the imputation stage are valuable in the EHR context because missingness and measurement interval irregularity are very common due to data being collected without a pre-specified protocol or research question in mind.[33,54,100]

While CRLI represents the latest advance in crisp IMVTS deep clustering, it has thus far only been applied in the context of acute respiratory failure in the ICU. Importantly, this study by Wu et al. did not utilize the imputation ability of CRLI and relied instead on traditional methods like forward, backward, and median filling. Furthermore, they did not address our biomedical area of interest, chronic disease treatment response.[32] Thus, to address this biomedical interest and do so in a way that would maximize the multivariate ability of CRLI, we chose to explore response trajectories of GLP-1 RA medications, a prime example of a drug class with diverse, multiorgan effects that are captured in the EHR during routine clinic appointments. While GLP-1 RA medication use in *All of Us* and other RWD sources has been explored, there has yet to be an investigation of longitudinal treatment response subgroups that utilizes IMVTS deep clustering.

3.2. Background and significance

3.2.1. Chronic disease responder endpoints in clinical trials

Chronic disease progression and treatment response can seldom be described comprehensively by a single outcome measure or biomarker. For example, measures of disease activity in rheumatoid arthritis include visual analogue scale pain assessment, swollen joint count, C-reactive protein level, and patient global evaluation.[101] Similarly, T2D therapy goals can include achieving control on various glycemic and metabolic targets (HbA1c, weight gain, SBP, LDL-cholesterol).[102] As such, composite responder endpoints are utilized in clinical trials in order to dichotomize patients into responders and non-responders based on multiple measures, such as in clinical areas like oncology, rheumatology, endocrinology, and others.[102–104] However, creating binary and/or composite outcomes at a single predetermined point in time muddies the importance of individual measures and discards the longitudinal trends intrinsic to treatment response.[105] Improvement in one individual measure but deterioration in another would not be apparent from an aggregate measure alone. Patients with chronic diseases exhibit significant between-patient and within-patient variability over the course of disease progression and treatment response. A composite measure representative of multiple dimensions of disease activity or outcomes may suggest that individual components are changing in parallel, though that would be a problematic assumption. At worst, treatment guidelines based on composite score benchmarks could lead to mistreatment or overtreatment. This is because benchmarks (1) can have different meanings in different patients and (2) do not illuminate which dimension of disease should be prioritized for treatment. A single multidimensional measure may be appropriate for clinical trial design due to improvement of signal-to-noise ratio and elimination of outlier effects. But for the same reasons, its utility does not translate to monitoring individual patients in routine clinical care who may experience a non-typical disease course.[101,106]

3.2.2. Real-world data: NIH *All of Us*

Real-world data, and EHRs in particular, capture this routine clinical care in structured (diagnoses, prescriptions, lab tests) and unstructured (clinical notes, imaging) formats.[47] While not a replacement for clinical trials, these datasets can be utilized to assess the comparative effectiveness of medications in practice and survey more heterogeneous populations of patients than trial eligibility guidelines may allow for.[107] In particular, previous studies have shown that patients enrolled in randomized controlled trials may not be representative of chronic disease patient populations. Therefore, investigation of these real-world datasets may identify formerly unknown patient subpopulations in need of novel therapeutic interventions.[96] Internal validity of trial results can benefit from stringent exclusion criteria that results in rejection of patients with complex disease. However, this may limit external validity of those trials, and those excluded patients may ultimately be subject to prescribing practices that are informed by those trials.[108]

One such dataset is the *All of Us* Research Program, an open-source, NIH-funded longitudinal cohort study which aims to collect comprehensive health data of at least one million participants from across the United States. Started in 2016, *All of Us* includes several biomedical data types, including EHR, genomics, physical measurements, wearables, and surveys.[109,110] Assessing treatment response in the context of *All of Us* real-world data, as opposed to carefully designed and curated randomized controlled trials, is more representative of the multifaceted and heterogeneous nature of routine clinical care. Approaches that analyze multiple disease-relevant routinely collected clinical biomarkers individually and simultaneously, while considering the entire longitudinal trajectory after initiation of treatment, present the opportunity to contribute to personalized medicine approaches in the present era of massive, multimodal real-world datasets.

3.2.3. Uncovering longitudinal treatment response subgroups

Analysis of longitudinal biomarker data across multiple facets of health can yield a richer depiction of a participant's changing disease status and response to treatment. Time series data available through the EHR can capture trends over time. For example, participants may be "fast" or "slow" responders to treatment, may improve across many dimensions or just a few, and may experience temporary or lasting change. Beyond biomarkers, other measures, like vital signs and physical measurements, can also be tracked over time to contribute to this bigger picture of participant health.

Analysis of longitudinal clinical measures, particularly in a retrospective context, is complicated by variable time frames, inconsistent time intervals, and overall missingness.[47] CRLI, and other recently developed deep learning-based time series clustering methods, tackle these issues by simultaneously performing missing value imputation and clustering across variables.[33,54] We propose that these tools can be used to identify participant subgroups which would not be represented in dichotomized response outcomes, as is typical of clinical trials. Further, this approach reflects the reality of typical clinical care, where gold standard disease severity indices may not be regularly assessed as they are in prospective trials. An analysis built on commonly used labs and measurements is ripe for translation to the clinical setting.

3.2.4. Varied effects of GLP-1

Glucagon-like peptide-1 receptor agonists (GLP-1 RAs) are increasingly commonly prescribed antidiabetic medications. GLP-1 RAs have been shown to lower blood glucose, reduce weight, ameliorate hypertension, and preserve renal function. This multiplicity of targets is reflected in composite endpoints used in GLP-1 RA trials.[111] Importantly, these specific effects are captured in the EHR by the routinely collected measures like HbA1c, BMI, systolic and diastolic blood pressure (SBP, DBP), and Glomerular Filtration Rate (GFR). These measures encompass both

(1) outcomes the GLP-1 medications were designed to achieve and (2) tests that are ordered with some regularity for T2D patients.

By applying multivariate trajectory clustering methods to EHR-derived clinical time series from *All of Us*, we sought to identify GLP-1 RA treatment response subgroups and characterize these subgroups based on their longitudinal patterns and associations with baseline characteristics and outcomes of interest. We hypothesized that patient subgroups would exhibit variable (optimal vs. suboptimal) multivariate GLP-1 RA response trajectories and that these trajectories would be associated with risk factors that could guide diagnostic practices in the future. This work provides GLP-1 RA-specific insights and lays a foundation to be extended to other chronic diseases, data modalities, and longitudinal, real-world biomedical datasets.

3.3. Related work

Our methodological approach borrowed principles from (1) statistical modeling of multivariate treatment response in clinical trials, (2) deep clustering of diabetes progression trajectories from the longitudinal EHR, and (3) evidence generation of GLP-1 efficacy from real-world data sources. A selection of related works representative of these areas is laid out here.

3.3.1. Modeling treatment response

Modeling multivariate treatment response is a relatively novel concept; until recently, univariate analyses were more common. Previous univariate works have applied mixture modeling to observational data to identify longitudinal subgroups of response in conditions including major depressive disorder (MDD), psoriasis, and obsessive-compulsive disorder (OCD).[112–114] Recently, Parodis et al. applied growth mixture modeling to multivariate disease activity indices and patient-reported outcomes from clinical trial data to identify 4 latent classes of systemic lupus erythematosus (SLE) evolution following the initiation of belimumab therapy.[115]

Real-world data, and EHRs in particular, capture routine clinical care in heterogeneous populations that observational cohort and clinical trial data, by design, may be unable to.[47,96,107,108] However, longitudinal EHR are often characterized by inherent data issues like missingness, irregular spacing of measurements, and variable time series lengths.[87,97] Fortunately, novel methods, including those based on deep learning, have been formulated to accommodate these issues.

3.3.2. Insights derived from longitudinal EHR data

Deep learning has been increasingly applied to EHR data for clinical informatics tasks like information extraction, outcome prediction, phenotyping, and deidentification.[98] In particular, representation learning is a form of deep learning that is popular for EHR analysis because it can

identify latent patterns and structures within unlabeled data. Learned representations can be used for downstream analyses like patient similarity analysis, cohort subphenotyping, predictive model-building, and data visualization.[97] One such application is deep clustering of multivariate time series.[28,69]

3.3.2.1. Trajectories of diabetes progression

Type 2 Diabetes is a suitable biomedical target for longitudinal subtyping approaches in the EHR, including multivariate time series clustering, because it (1) is a chronic disease with increasing prevalence, (2) is associated with long-term, multiorgan complications like nephropathy and retinopathy, and (3) has multiple goals of therapy, including HbA1c reduction and less or no weight gain.[102,116] To date, a number of groups have used deep clustering methods to characterize diabetes progression trajectories.

Carr et al. developed a recurrent neural network autoencoder to cluster EHR data using reconstruction, outcome, and clustering losses and proposed flexible balancing of those losses. They proposed that combining the standard reconstruction loss with a time-to-event (outcome) loss could discover clusters of patients with both different trajectories and outcomes. The final list of features they clustered on included 286 primary diagnosis codes, 351 secondary diagnosis codes, 175 procedure codes, 122 medication types, and 55 laboratory values. They trained three versions of their model on a cohort of 29,299 diabetes patients: (1) without outcome loss, (2) without reconstruction loss, and (3) with combined reconstruction and outcome loss (see Section 2.4.1 for a discussion on types of losses used to learn model parameters in deep clustering). Their outcome of interest was time to first cardiac event. They trained the models multiple times to find clusters from k = 2 to k = 7 and demonstrated that the cluster Kaplan-Meier (KM) time-to-event curves change significantly depending on which loss functions are used. However, beyond

plotting KM curves for k = 3 and k = 5, they did not provide clinical characterization of the clusters.[1] Similar outcome-guided time series clustering methods are discussed in Section 6.4.2.2.

Manzini et al. utilized a kernelized autoencoder framework to identify seven trajectories of type 2 diabetes progression using eleven routinely collected EHR measures, including glycated hemoglobin (HbA1c), body mass index (BMI), diastolic and systolic blood pressure (DBP and SBP), lipid profile, and renal function, across a five-year timespan. They then described clinical characteristics (common comorbidities, antidiabetic treatments) in each cluster to identify differential features that typified each clusters' phenotype. Based on this differentiability, each cluster was given a representative name: Neuropathic, Hypercholesteraemic, Multiple Complications, Vascular Disease, Hypertensive, Retinopathy, and Metabolic. For example, while the Metabolic and Multiple Complications Clusters had characteristics in common (age at diagnosis, diabetes duration, high HbA1c levels), the Metabolic Cluster had more intensive pharmacological treatment and a lower rate of patients developing comorbidities. The authors propose that longitudinal clinical phenotypic evolution profiles like these can inform personalized treatments for T2D patients.[2]

To our knowledge, one-stage deep MVTS clustering methods have yet to be employed with respect to antidiabetic treatment response using data from real-world sources. This work is especially relevant in the context of the availability of diverse treatment options with varied mechanisms of action which can potentially optimize different treatment outcomes for patients. We constrain our focus to GLP-1 receptor agonists, a relatively new class of antihyperglycemic drugs, which are being increasingly utilized due to their additional positive effects on cardiovascular, renal, and metabolic health.[3]

3.3.2.2. GLP-1 analyses

A number of groups have analyzed longitudinal EHR data to characterize GLP-1 RA (1) effectiveness and risks, (2) utilization, discontinuation, and reinitiation patterns, and (3) treatment responsiveness. In their exploration of 7,239,854 person-years worth of EHR data from the US Department of Veterans Affairs, Xie et al. systematically assessed associations between GLP-1 RA use and 175 diverse health outcomes across 12 diagnostic categories. They showed that, compared to usual care (non-GLP-1 RA antihyperglycemic regimen), GLP-1 RA use was associated with broad multidimensional effects that reinforce and extend those described in the existing literature.[3] In their review of real-world evidence, Thomsen et al. noted that the effectiveness of GLP-1 RA for weight loss is supported, but not at the levels seen in clinical trials, likely because of reduced adherence and persistence. This discontinuation is frequently caused by adverse events, especially gastrointestinal disturbances, and high medication costs. The authors stress that future research should assess long-term clinical impacts and effects of treatment discontinuation.[4] Rodriguez et al. explored this aspect in their analysis of a subset of Truveta Data, linked EHR data from a collective of 30 US healthcare systems. Consistent with other analyses, they found that the majority of patients who newly initiated GLP-1 RA treatment discontinued within 1 year.[6]

In another analysis of Truveta data, Rodriguez et al. compared the efficacy of semaglutide and tirzepatide for achieving clinically meaningful weight loss in a propensity-matched cohort, finding that tirzepatide was the more efficacious medication and that patients without T2D experienced larger weight reductions than those with T2D.[5] Zhu et al. developed machine learning models based on common clinical factors recorded in the EHR (demographics, lab values, medication use, comorbidities) to predict GLP-1 RA and DPP4i responsiveness (defined as 0.5% HbA1c reduction) in a cohort of almost 8,000 patients seen at Vanderbilt University Medical Center.[7]

Cardoso et al. also used routine clinical features in the UK EHR to design a Bayesian methods-based treatment selection algorithm capable of predicting differences in glycemic outcomes for GLP-1 RA and SGLT2i therapies.[117]

Finally, several studies have explored similar GLP-1 RA themes within the *All of Us* dataset. Devineni et al. characterized SGLT2i and GLP-1 RA prescribing trends in patients with diabetes by cardiovascular risk and sociodemographic factors.[8] Mariam-Smith et al. used logistic regression to investigate the impact of genetic variation in the *NBEA* gene on GLP-1 RA weight loss.[9] Salvatore et al. used a PheWAS approach to compare novel diagnoses following GLP-1 RA, SGLT2i, and DPP4i prescription in T2D patients, highlighting semaglutide's unique downstream association profile.[118] Mayer and Fontelo (1) reported BMI, weight, and HbA1c changes in semaglutide users, (2) assessed differences in outcomes by route of administration (injectable vs. oral), and (3) examined development of common adverse events after starting the medication.[10]

Taken together, these studies of the longitudinal EHR demonstrate the potential of routinely collected health data to inform precision medicine approaches to T2D prescribing practices. We designed a methodological approach to use one-stage deep MVTS clustering to extract latent longitudinal subgroups of GLP-1 RA treatment response as captured by routinely collected measures in the *All of Us* EHR.

3.4. Methods

3.4.1. The *All of Us* Research Program: EHR domain

The *All of Us* research protocol, recruitment structure, timeline, scientific goals, and data access, harmonization, quality, utility, missingness, and diversity have been described in detail elsewhere.[109,119–124] Our work utilized the EHR domain of the *All of Us* Controlled Tier Dataset v7 Curated Data Repository (CDR), released in September 2024.[125] The v7 CDR, which has a data cutoff date of 07/01/2022, includes health data for 413,457 participants total, of which 287,012 had EHR data available.[126] EHR data are organized into five domains: condition occurrence, observation, drug exposure, procedure, and measurement. Within those domains, data types available include demographics, visits, diagnoses, procedures, medications, laboratory tests, and vital signs. EHR data collected by the program are transformed into the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) structure.[127–129] This transformation process includes validation checks, characterization checks, data cleaning, and suppression of certain tables and fields to ensure privacy protection.[130]

3.4.2. Cohort selection

3.4.2.1. GLP-1 RA prescription history

Using the *All of Us* Researcher Workbench Cohort Builder, we first narrowed our cohort to only those with any history of GLP-1 RA prescription (n = 10,367). We specified this with ATC code A10BJ (OMOP concept ID 1123618), “Glucagon-like peptide-1 (glp-1) analogues”, which included the medications and participants counts shown in Table 7.[131]

GLP-1 analogue	Count
albiglutide	109
dulaglutide	4,556
exenatide	1,347
liraglutide	4,019
lixisenatide	61
semaglutide	3,739

Table 7. Counts per GLP-1 drug

We used the Dataset Builder to generate a SQL query for the “drug” domain, which built a dataframe containing fields from the “drug_exposure” OMOP table and from the source

vocabulary (in this case, RxNorm).[132] Entries in each field can contain varying levels of information. For example, "standard_concept_name" can contain dosage, route of administration, combination therapy medications, and brand name (e.g., "0.25 MG, 0.5 MG Dose 1.5 ML semaglutide 1.34 MG/ML Pen Injector"), but can also be as simple as "semaglutide".[133] The RxNorm structure revolves around the Concept Unique Identifier (RxCUI), which is used to link related entities, like drug names, ingredients, and classes.[134] To focus our analysis, we created a new field, called "drug_simple_name", by extracting the GLP-1 RA drug name only from "standard_concept_name" (as shown in Table 7).

3.4.2.2. Drug exposure duration

Some fields have higher missingness rates than others. For example, while "drug_exposure_start_datetime", known henceforth as "drug_start", had no missing values, "drug_exposure_end_datetime", hereafter "drug_end", was missing for 32% of prescriptions in our cohort. First, we used "drug_start" to characterize the overall per-year GLP-1 prescription volume in *All of Us*. Then, in the absence of reliable prescription lengths and adherence information, we sought to approximate the length of time for which each participant was consecutively prescribed, and therefore exposed to, their first recorded GLP-1 medication. We selected only each participant's earliest "drug_start" for each "drug_simple_name". We then selected the latest recorded "drug_start" and "drug_end" for that "drug_simple_name" for each participant. For whichever of these two was the latest, we took the difference between it and the participant's "drug_start", and called this "drug_duration" (Figure 32). Participants with overlapping prescription windows were removed.[d] Finally, we selected only the earliest prescribed "drug_simple_name", based on "drug_start", for each participant. We included only the single

[d] If a participant was at any point prescribed different GLP-1 drugs simultaneously (overlapping EHR prescription records), they were excluded from the cohort because our modeling approach could not account for concurrent medication use.

earliest GLP-1 medication per participant in our dataset so that (1) the same participant was not included multiple times in our analysis and (2) each participant's outcome measures reflected GLP-1 naiveness. However, this does not rule out the possibility that participants had taken GLP-1 medications even earlier which were just not recorded or harmonized in the *All of Us* EHR. Having computed the start date, end date, and drug_duration for this earliest prescribed GLP-1 medication (per participant), we assessed how many times each of our outcome measures of interest were recorded leading up to and while each participant was on the medication.

3.4.2.3. Routinely collected longitudinal measures

We used the Dataset Builder to generate dataframes for participant demographics ("person" domain) and our five outcome measures of choice ("measurement" domain, Table 8) which we merged with our GLP-1 duration dataframe. Using participant sex, age, and serum creatinine, we calculated eGFR using the 2021 CKD-EPI Creatinine Equation. This equation, which is non-race-based, is the recommended method for estimating GFR in adults from the National Kidney Foundation.[135,136]

Measure	**OMOP ID**	**LOINC code**
Body mass index (BMI) [Ratio]	3038553	39156-5
Systolic blood pressure	3004249	8480-6
Diastolic blood pressure	3012888	8462-4
Hemoglobin A1c/Hemoglobin.total in Blood	3004410	4548-4
Creatinine [Mass/volume] in Serum or Plasma	3016723	2160-0

Table 8. Longitudinal measure codes

American Diabetes Association (ADA) monitoring guidelines recommend assessing HbA1c a minimum of two to four times per year, depending on treatment goals.[137] We anticipated that, for most participants, time windows of (1) 6 months pre-treatment start would capture at least one set of measurements, to serve as a baseline, and (2) 2 years post-treatment would capture at least four sets of measurements, to capture longitudinal therapeutic effects. Thus, using a base unit of 90 days to represent 3 months, we selected only those participants who had a

"drug_duration" greater than or equal to 720 days. Then, for each outcome measure, we removed values that were extreme outliers (z-scores > 10) across the entire cohort. Finally, we assessed measurement availability (availability of the measurement data records for a given patient at a given point in time) of our five measures such that we had at least 1 measurement of each in the 180 days prior to "drug_start" and 2 measurements of each measure in the 720 days after. This left us with a final cohort size of 336 (Figure 32).

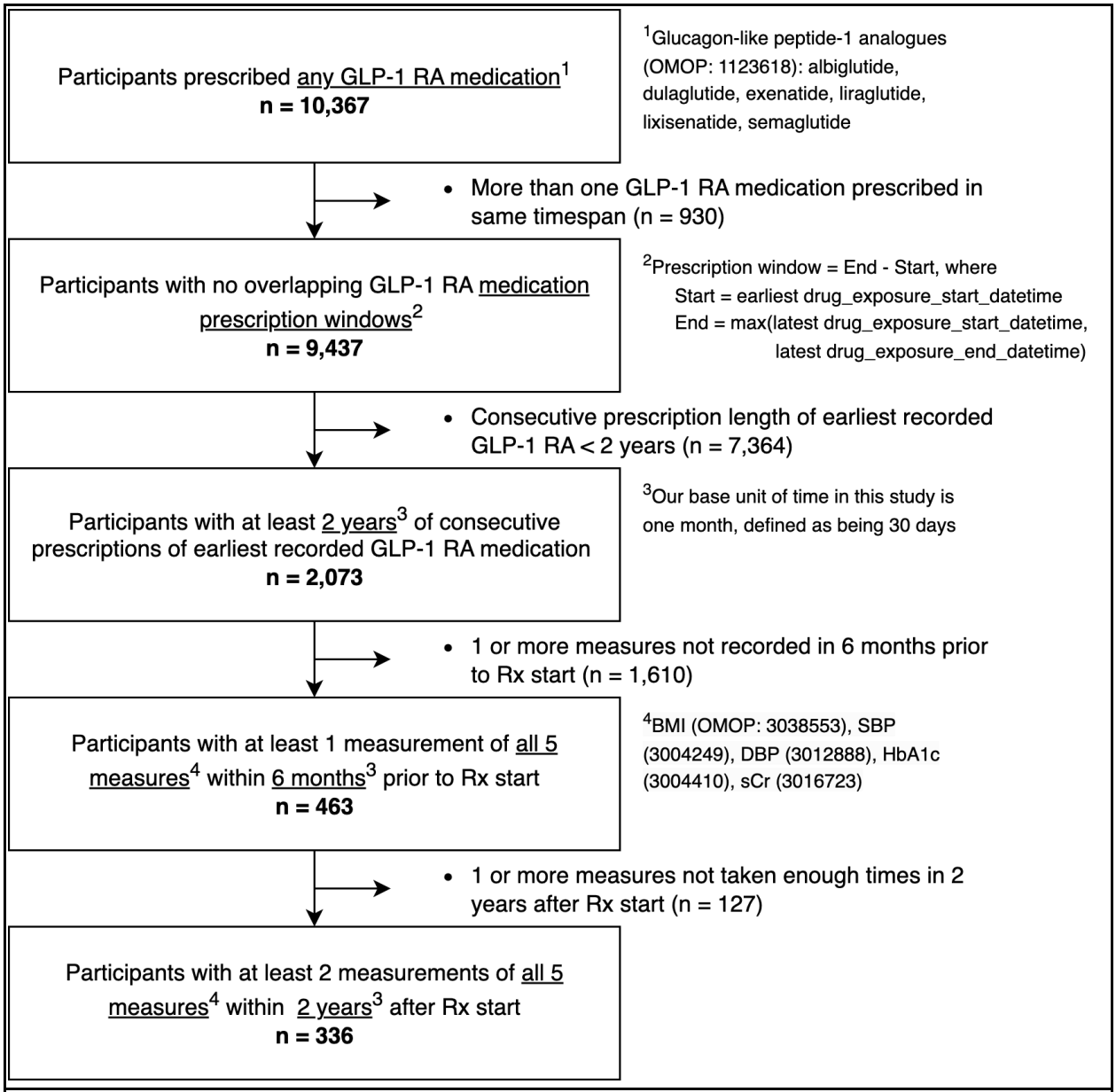


Figure 32. GLP-1 RA cohort selection diagram

3.4.3. Data preprocessing

3.4.3.1. Time series regularization

CRLI is unable to accommodate time series of varying length, meaning that each input time series must have the same number of data points, though any of those data points can be missing (see Section 6.4.1.4 for further discussion). Thus, to regularize each participant time series to a fixed length with equal-sized segments, we applied a modified version of Piecewise Aggregate Approximation (PAA). This method divides our 900 day total window length (180 days before treatment initiation, 720 days after) into 10 equal-sized segments of 90 days (3 months) each and calculates the mean for any variable which is measured more than once in the same segment (see Section 3.6.2.2. for discussion on window size).[46,138,139] Though the original implementation of PAA was intended for denoising highly sampled time series, we slightly modified it to handle sparse time series as well (Figure 33).[140] The *All of Us* Data User Code of Conduct prohibits dissemination of participant-level data, thus Figure 33 shows randomly generated synthetic data to demonstrate PAA behavior.[141]

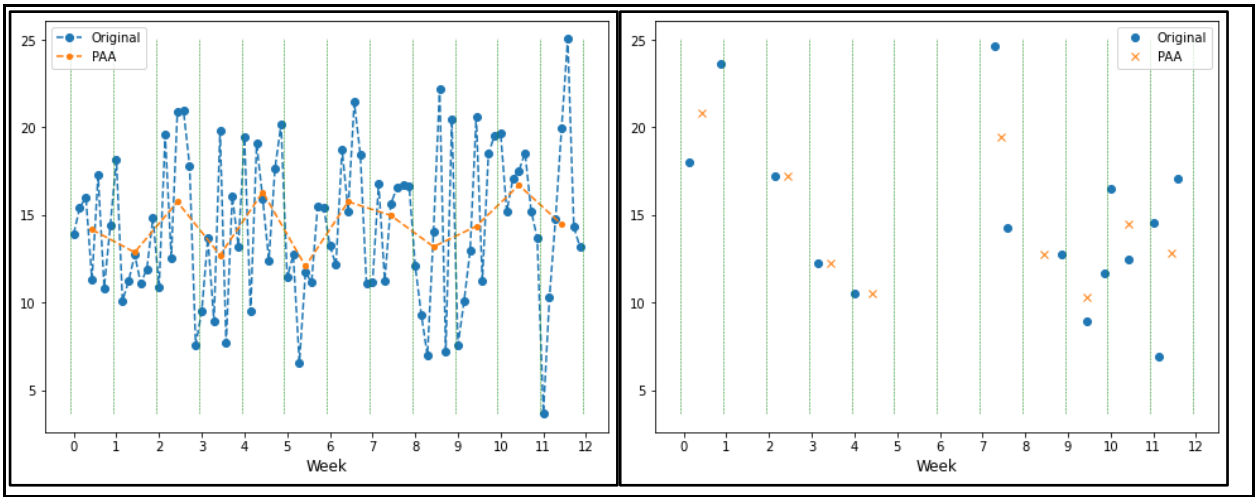


Figure 33. Original (denoising) vs. modified (sparse) PAA behavior[142]

3.4.3.2. Normalization relative to baseline

Before applying PAA, for each measure, we had taken each participant's last measurement before "drug_start" to be their pre-treatment baseline. After applying PAA, following de Jong et al., all measures were normalized relative to baseline by (i) subtracting the baseline mean across all patients and (ii) dividing by the baseline standard deviation across all patients.[54]

3.4.4. Clustering Representation Learning on Incomplete time-series data (CRLI)

CRLI is among the recently developed deep learning methods which simultaneously perform missing value imputation and clustering across MVTS (Figure 30). We used the PyPOTS implementation of CRLI. PyPOTS is a Python toolbox which supports imputation, classification, clustering, forecasting, and anomaly detection tasks on multivariate Partially-Observed Time Series with missing values.[143]

The CRLI hyperparameter search space described by Ma et al. (2021) is summarized in Table 9. Further details can be found in the original CRLI paper, the corresponding GitHub repo, and the PyPOTS GitHub repo.[33,144,145] Due to computational constraints, we used the same default hyperparameters (underlined in Table 9) as in the original paper. We set the PyPOTS early stopping mechanism (patience), which tracks performance based on generator training loss, to 20 epochs out of a total of 500 epochs.[146] Since we did not have ground truth cluster labels to evaluate against, we trained CRLI on the full dataset and predicted cluster membership of the same full dataset, rather than splitting our data into train and test sets.[147]

Hyperparameter	CRLI repo name	PyPOTS repo name	Search space
Coefficient of the soft K-means objective	lambda_kmeans	lambda_kmeans	{1e-3,1e-6,1e-9}
# of units of each layer in the encoder	G_hiddensize	rnn_hidden_size	{50,100,150}
# of layers of the encoder	G_layer	n_generator_layers	{1,2,3}

Table 9. CRLI hyperparameters

3.4.5. Clustering validation indices

Clustering validation indices (CVI) are used to evaluate unsupervised clustering results. The creators of CRLI evaluated its performance with external CVI (Rand Index, Normalized Mutual Information, Cluster Purity, Cluster Accuracy) on datasets where the ground truth labels were known.[27,33,66] In this work, we utilized internal CVI to select the optimal k (number of clusters), from a range of k values, since we do not have ground truth subgroups to compare against. Internal CVI are designed to evaluate the goodness of a clustering result without any outside information based on (1) compactness, which measures how closely related members within the same cluster are, and (2) separation, which measures how distinct one cluster is from all others.[84] However, no CVI is supremely performant across every clustering context, so utilizing multiple is recommended. Drawing from related work and comparative studies, we specifically aimed to maximize Calinski-Harabasz and Silhouette indices and minimize Davies-Bouldin and S_Dbw validity indices (Table 5).[68,84–87,89,148,149] In our lineplots showing internal CVI performance for each k, we highlighted in red the k for which each index was optimized (Figure 36).

For more discussion on the differences between external and internal CVI, see Section 2.4.4.

3.4.6. Coding workflow

While not identical to Aim 1, our Aim 2 workflow writeup is more refined and can be found in Section 4.4.9.

3.5. Results

3.5.1. Yearly GLP-1 prescription volume in *All of Us*

In 2021 alone, 27,334 GLP-1 prescriptions were written for 5,557 unique participants in *All of Us* (Figure 34). Though the complete prescribing trend is not reflected in Figure 34 since CDR v7 has a data cutoff date of 07/01/2022, we would expect to see the same exponential trendline in future data releases. Semaglutide (first approved for medical use in the US in 2017) and dulaglutide (2014) made up the bulk of written GLP-1 RA prescriptions in recent years, with liraglutide's (2010) portion decreasing.

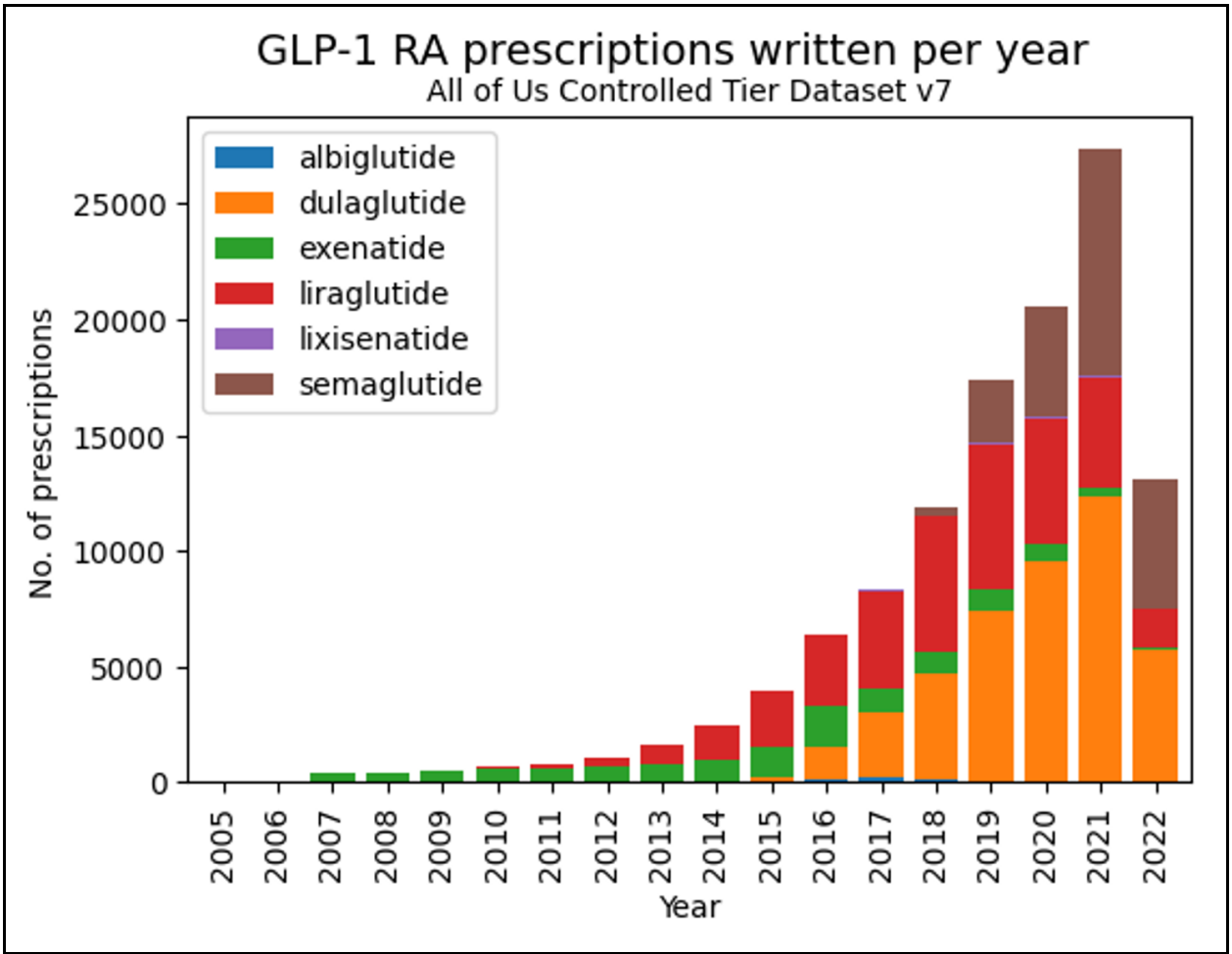
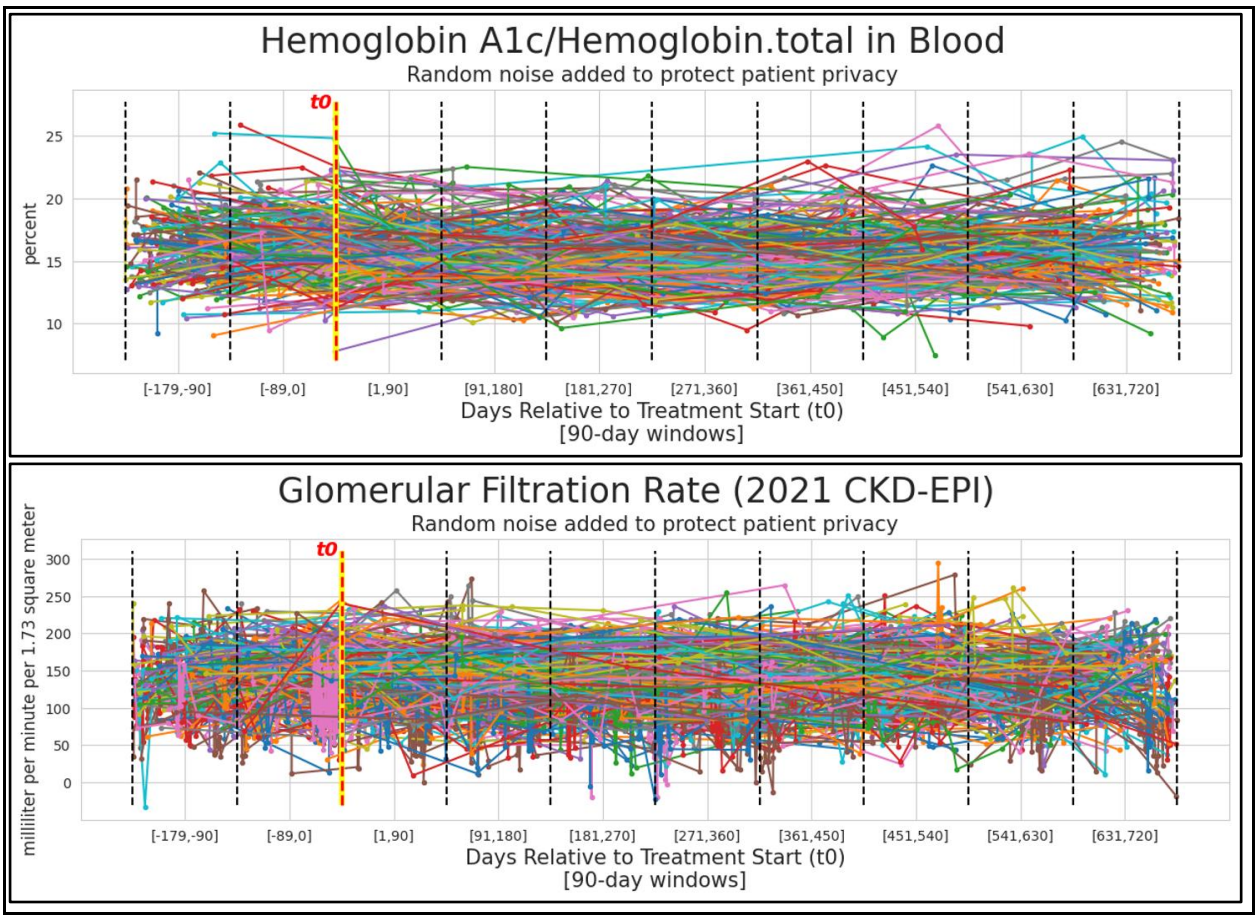


Figure 34. Yearly GLP-1 RA prescription volume

3.5.2. Individual GLP-1 RA outcome trajectories

A visualization of our methodological approach, Figure 35 shows all 336 participant trajectories (with random noise added[e] to comply with *All of Us* data dissemination policy) for three outcome measures.[141] These measurements were taken in the range (-6 months, +2 years] relative to treatment start (t0). Dashed lines demarcate the ten 90-day windows to which PAA was applied to regularize time series lengths before applying CRLI.



[e] Each value was perturbed by Gaussian noise with mean & SD matched to those of the original variable

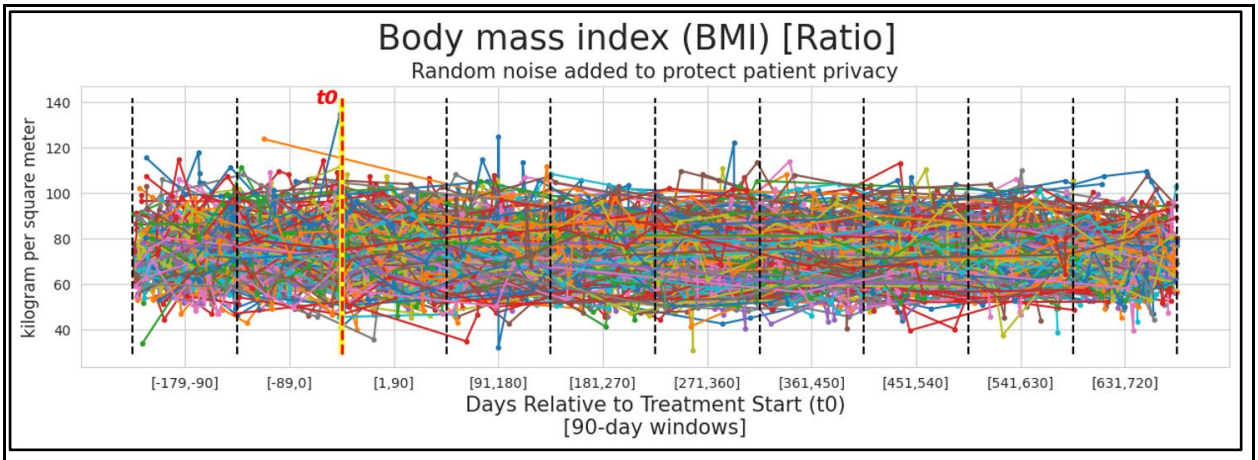


Figure 35. Individual outcome measure trajectories

3.5.3. Optimal cluster number selection

We concluded that 5 clusters was optimal based on (i) reaching the elbow of the S_Dbw index and (ii) relative minimization of the Davies-Bouldin score. However, we also generated a 2-cluster result based on (i) maximization of silhouette score and Calinski-Harabasz score, (ii) relative minimization of Davies-Bouldin score and (iii) greater likelihood of clinical interpretability. Figure 36 shows calculations of each index for CRLI results specified on 2 to 8 clusters. Red points indicate the cluster number for which the index achieved the best performance.

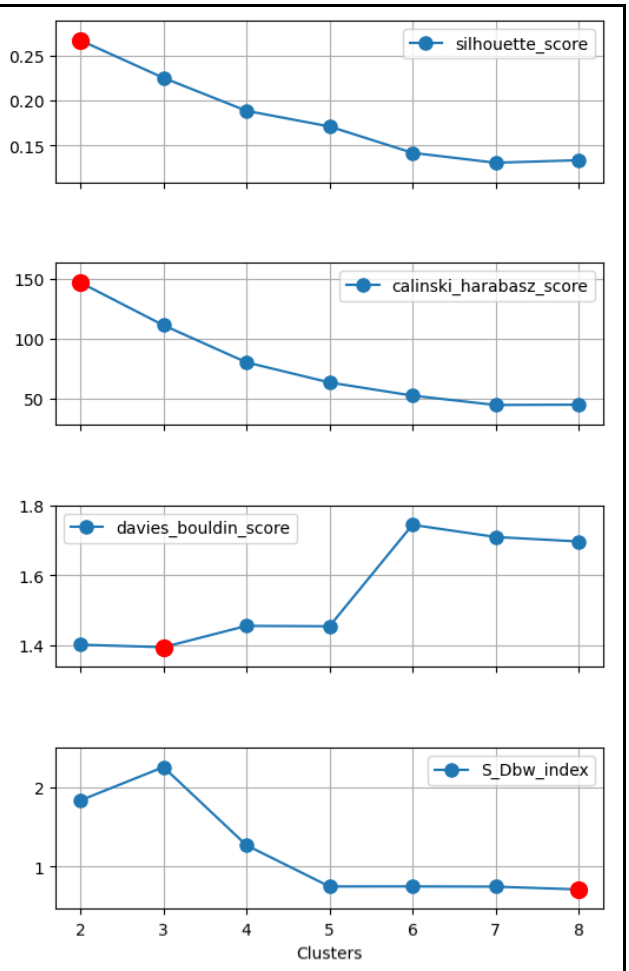


Figure 36. Internal CVI for GLP-1 RA outcome clusters

3.5.4. Trajectory cluster mean lineplots

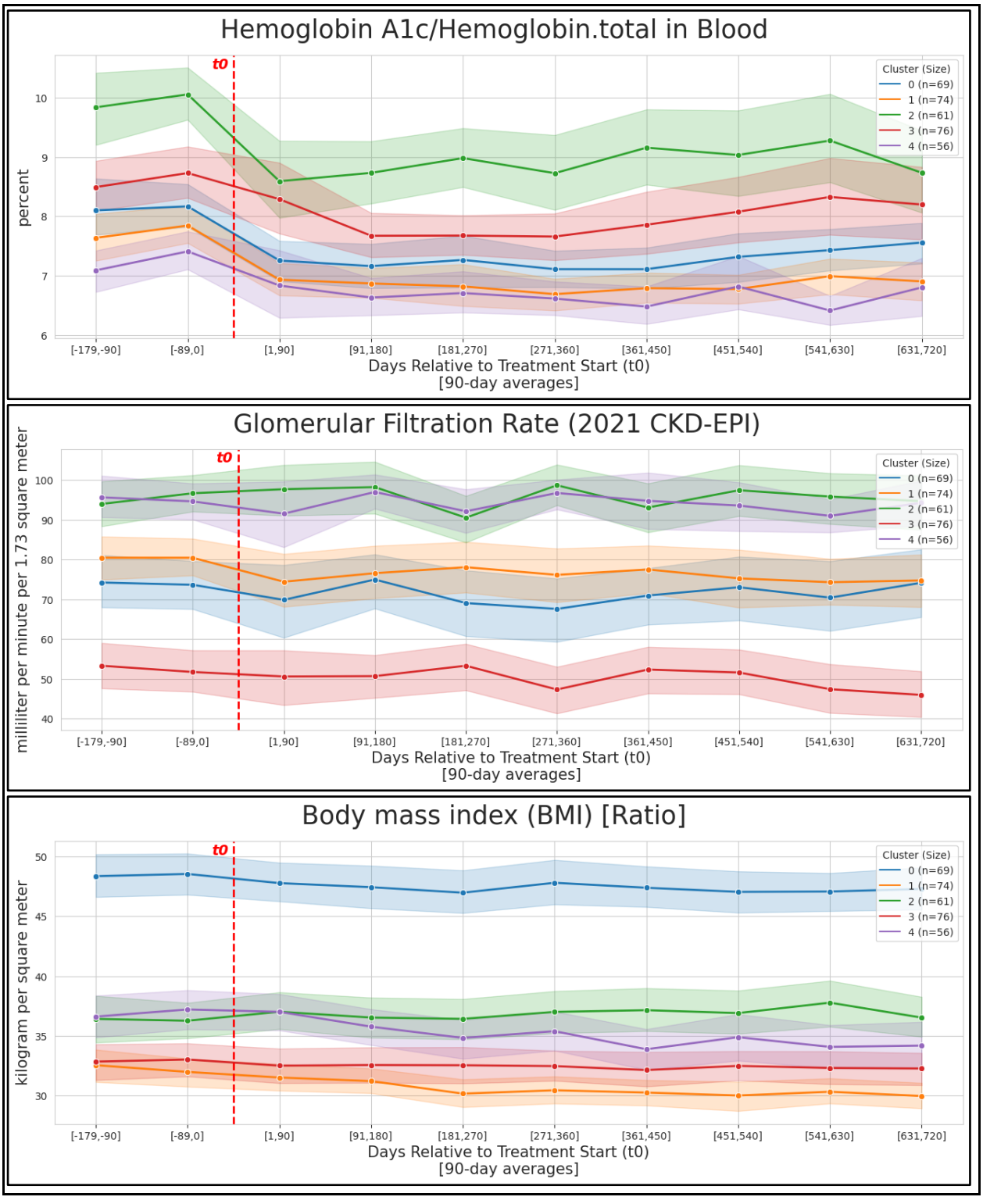
3.5.4.1. Five clusters

All clusters experienced a mean drop in HbA1c, with Cluster 2 (green, n=61) starting the highest and having the steepest decrease. Clusters 2 and 4 (purple, n=56) fell within the normal eGFR range (>90 mL/min/1.73m^2), Clusters 0 (blue, n=69) and 1 (orange, n=74) fell within the early-stage kidney disease range (60-89), and Cluster 3 (red, n=76) fell in the kidney disease range (15-59). Interestingly, BMI trajectories were relatively stable[f] across clusters, though Cluster 0 fell into Class 3 (severe) obesity. Cluster 1 was the only group to fall within normal blood pressure ranges; Clusters 0 and 2 were elevated and Clusters 3 and 4 would be considered stage 1 hypertension. These findings are summarized in Table 10 and mean lineplots are shown in Figure 37.[150–152] Our clinical characterization of these clusters must be interpreted critically, given the cohort selection limitations discussed in Section 3.6.2.1.

Cluster	HbA1c (% change from pre- to post-t0)	eGFR (clinically relevant ranges)[g]	BMI (obesity class)[h]	SBP & DBP[i]
0 (blue)	-11.2	60 - 89	Class 3 (40+)	Elevated (SBP: 120-129, DBP: <80)
1 (orange)	-11.6	60 - 89	Class 1 (30-35)	Normal (SBP: <120, DBP: <80)
2 (green)	-14.5	> 90	Class 2 (35-40)	Elevated
3 (red)	-5.1	15 - 59	Class 1	Stage 1 hypertension (SBP: 130-139, DBP: 80-89)
4 (purple)	-7.8	> 90	Class 2	Stage 1 hypertension

Table 10. Summary of GLP-1 RA trajectory trends

[f] Absolute value of percent change did not exceed 5 for any cluster between any timepoints
[g] Kidney disease-relevant eGFR ranges are from the National Kidney Foundation[150]
[h] Obesity classes are from the Centers for Disease Control and Prevention[151]
[i] Blood pressure categories are from the American Heart Association[152]


Hemoglobin A1c/Hemoglobin.total in Blood
t0
percent
Cluster (Size)
0 (n=69)
1 (n=74)
2 (n=61)
3 (n=76)
4 (n=56)
[-179,-90]
[-89,0]
[1,90]
[91,180]
[181,270]
[271,360]
[361,450]
[451,540]
[541,630]
[631,720]
Days Relative to Treatment Start (t0)
[90-day averages]
Glomerular Filtration Rate (2021 CKD-EPI)
t0
milliliter per minute per 1.73 square meter
Cluster (Size)
0 (n=69)
1 (n=74)
2 (n=61)
3 (n=76)
4 (n=56)
Days Relative to Treatment Start (t0)
[90-day averages]
Body mass index (BMI) [Ratio]
t0
kilogram per square meter
Cluster (Size)
0 (n=69)
1 (n=74)
2 (n=61)
3 (n=76)
4 (n=56)
Days Relative to Treatment Start (t0)
[90-day averages]

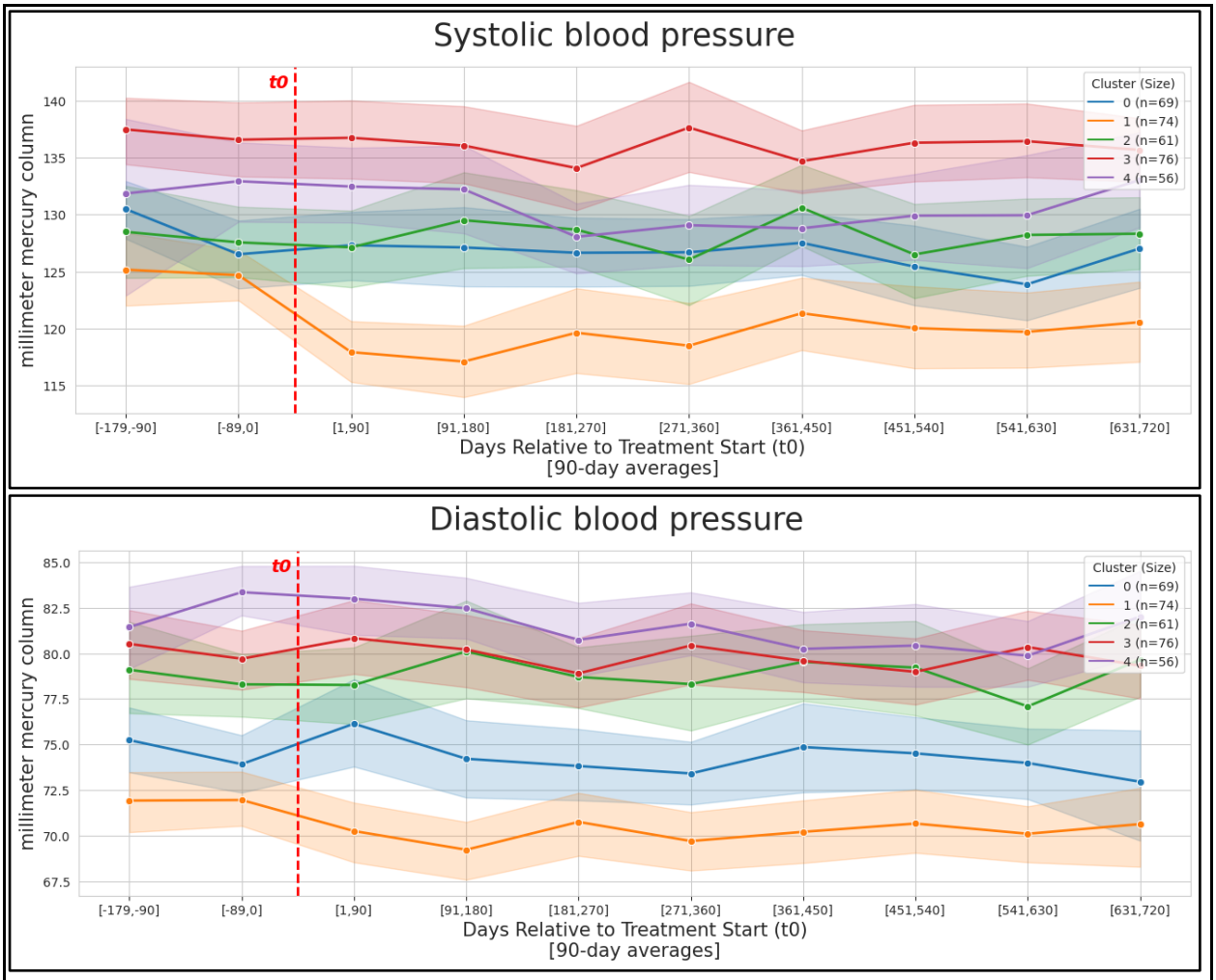
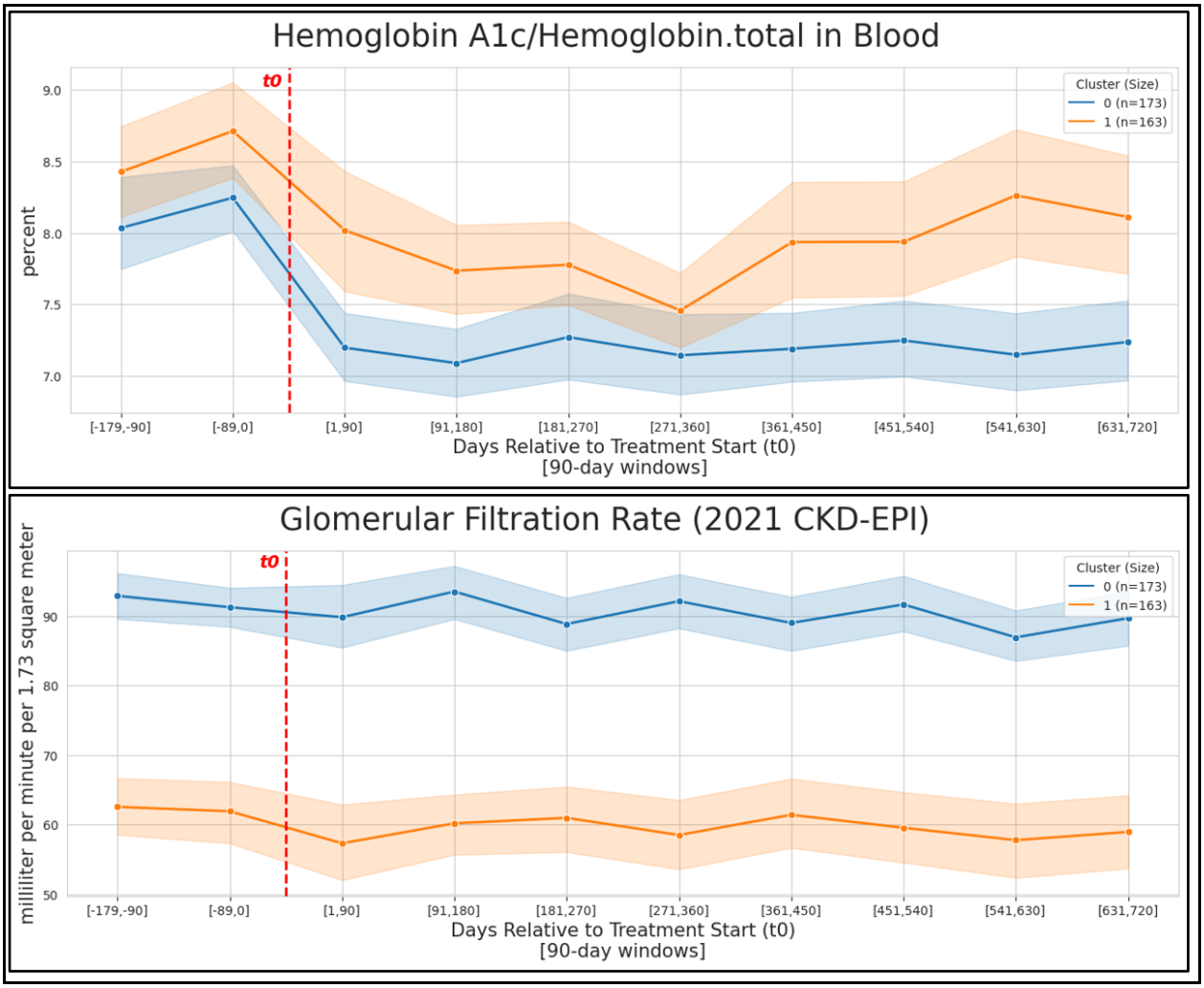


Figure 37. Multivariate trajectories of GLP-1 RA response (5 clusters)

3.5.4.2. Two clusters

Three of the total five outcome trajectories are shown in Figure 38. A decrease in average HbA1c after initiation of therapy was seen in both clusters. Interestingly, around the one year mark, Cluster 1 (orange, n = 163) HbA1c begins to rise while Cluster 0 (blue, n = 173) remains stable. Cluster 0 falls within the normal eGFR range while Cluster 1 can be characterized as at risk for kidney disease. Lastly, Cluster 0 BMI is within Class 1 obesity range whereas Cluster 1 is at the high end of Class 2. Though each cluster was characterized by higher or lower mean eGFR and BMI, no positive or negative trend post-treatment was seen in either. SBP and DBP (not shown) were similarly unremarkable.



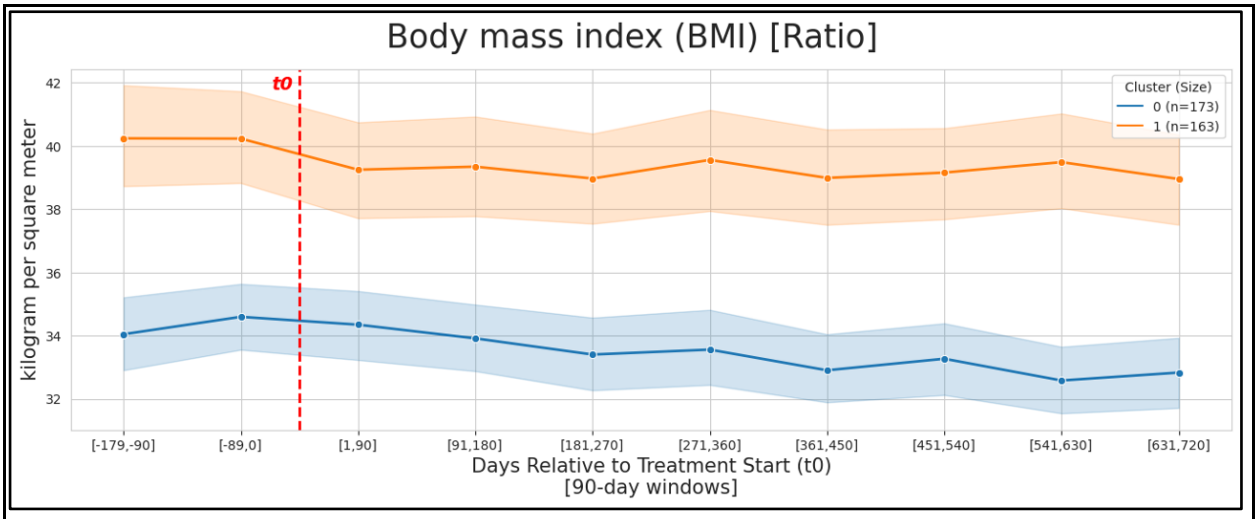


Figure 38. Multivariate trajectories of GLP-1 RA response (2 clusters)

3.6. Discussion

3.6.1. Conclusions from results

3.6.1.1. Informatics contributions

We have successfully deployed CRLI in a real-world data setting (electronic health record), showing that even short, sparse, irregularly sampled multivariate time series data can contain enough signal to identify distinct latent subgroups. In particular, this represents the first use of CRLI to characterize multiple longitudinal treatment response outcomes in the context of chronic disease. This complements work by others which have (1) applied deep clustering methods to study common diabetes progression trajectories, (2) applied latent class mixed modelling to define trajectories of treatment response in patients with other chronic diseases (e.g. psoriasis, lupus), and (3) applied CRLI specifically to characterize sepsis-induced acute respiratory failure in the ICU.[2,32,112,115] As has been described elsewhere, we demonstrated the application of PAA to temporal EHR data to regularize time series length to mimic the cadence recommended by routine monitoring guidelines.[46] Our overall approach was informed by components of pharmacoepidemiology, clinical trial subgroup analysis, and treatment pathway elucidation from observational data.[58,153,154] We also made contributions to the open-source PyPOTS repository, as mentioned in Section 4.6.1.1.

3.6.1.2. GLP-1 prescription landscape

GLP-1 prescription volume has been exponentially increasing for the past several years, with the bulk of that volume made up by dulaglutide and semaglutide since 2021. This finding is supported by other EHR analyses.[155] It will be crucial to continue to monitor these participants longitudinally as more data points become available in future data releases. To that end, the *All of Us* CDR v8 was made available in February 2025 and includes more than a year more of data (cutoff date of 10/1/2023).[125] Person-centered, longitudinal subtyping like that performed in this work lays the foundation for a more complete understanding of the patient experience over time.

3.6.1.3. Longitudinal GLP-1 response subtypes

CRLI identified distinct subgroups of multivariate response to GLP-1 treatment measured during the course of routine care. Beyond the quantitative separation of clusters evident from internal CVI and trajectory mean lineplot visualizations, we also found that each subgroup's longitudinal measures generally fell within clinically relevant and actionable ranges. These findings were hypothesis generating and motivated further phenotypic characterization of clusters. For example, we would expect Cluster 3 (Figure 37) to be enriched for kidney disease based on consistently low eGFR levels. Similarly, we would expect Clusters 3 and 4 to have hypertension-related comorbidities due to elevated systolic and diastolic blood pressure levels.

Interestingly, most subgroups generally had post-treatment longitudinal stability across outcome measures. This calls into question the added value of using time series clustering, at least in this context, as opposed to cross-sectional clustering. Richer phenotyping of clusters, as described below in Future Directions, may lead us to a firmer conclusion. If we were to use cluster membership as a predictor for future outcomes, it may be that simple aggregate measures, like quantiles or percent change from baseline, would be just as performant. The stability of BMI in particular casts doubt on the efficacy of these medications as antiobesity agents. However, it may be that BMI stability is itself a treatment success because it is better than weight gain. This motivates a parallel exploration of untreated T2D trajectories or a comparison with other antihyperglycemic medications, like SGLT2 inhibitors and DPP-4 inhibitors.[2,156]

3.6.2. Limitations

3.6.2.1. Cohort selection

Our "drug_start"- and "drug_end"-based medication sequencing logic is unlikely to have fully captured the true per-participant duration of drug exposure. We did not have reliable data regarding adherence, discontinuation, and adverse effects, though these have been

characterized in other datasets.[6,157,158] Other approaches to medication sequencing logic have additional stipulations, like requirement of repeat exposures to a given drug within a specified timespan.[109,154] We also did not have consistent information on dispensing, tracked via the "drug_type_concept_id" field, though by EHR analysis conducted by Gratzl et al., we can expect that around 70% of prescriptions were dispensed within 60 days.[130,155] We also did not consider the indication for which the medications were prescribed (diabetes vs. obesity), neither by EHR-based phenotyping nor by existence of the relevant diagnostic code in the patient's history.[10,159]

7,364 out of 9,437 (78%) participants were excluded due to having a consecutive prescription length of less than two years (Figure 32). This is in line with analyses which have found GLP-1 RA discontinuation rates up to 75% at 12 months.[6,157] Thus, while the proportion of included participants is unlikely to increase drastically in future *All of Us* data releases, the total number of included participants will rise as more individuals are enrolled in the study. Since CDR v7 was limited to EHR history before July 2022, we were not able to capture the introduction of tirzepatide, approved in the US in May 2022, into the prescribing landscape.[155]

Our final cohort was further limited by the strict measurement availability criteria. Because inclusion of three timepoints allows for non-linear trajectory estimation, we required a total minimum of three measurements of each outcome measure (1 before treatment start [t0], 2 after treatment start [t0]).[63] However, our final cohort represented only 3% of all participants ever prescribed a GLP-1 drug. This 3% is likely very different from the overall population–they could be (1) more chronically ill, requiring greater frequency of clinical visits, (2) of higher socioeconomic status, giving them greater leeway with insurance to remain on medications for longer than others, (3) be suffering from fewer adverse reactions (not reflected in our outcome measures) that would have otherwise caused discontinuation, and/or (4) tolerating the medication but not benefitting from it enough to discontinue within two years.[6]

Future application of MVTS clustering application in contexts like this will require careful consideration of how much cohort selection guidelines necessary for methods feasibility limit the final cohort from being representative of the actual population being treated.[97] Research design could include a thorough characterization of repeat measurement availability and longitudinal missingness before attempting to perform clustering.[160] Risks of subtyping in real-world data and cohort generalizability concerns are discussed further in Section 6.4.1.3.

3.6.2.2. Data preprocessing

Routine monitoring (measured every 3-6 months) of HbA1c is recommended for type 2 diabetics, but this frequency was not reflected in many participants' EHR. We set our PAA window size to 90 days to approximately capture this recommended cadence but it may be worthwhile to iterate through multiple other window sizes. Further, though we required 3 measurements total of each measure per participant, we did not stipulate where in the post-treatment window the measurements had to fall. Therefore, there could have been some participants with measurements bunched together and others with a wider spread. Lastly, after removing extreme outliers, we assumed that measurements without unit information were most likely the same as the standard unit for that lab. However, we could have applied more thorough unit harmonization strategies.[161,162]

3.6.2.3. Clustering

See Section 4.6.2.3 for discussion of limitations related to CRLI hyperparameter search space and internal CVI usage. Optimal cluster number selection in the absence of ground truth labels remains an open problem. Though a number of indices have been developed, these often give conflicting results (Figure 36).[66,84] Our quantitative assessment of cluster separation lacked a

complementary qualitative component of how well cluster trajectories represented actual clinical relevance and generalizability from the judgment of a medical domain expert.[32]

3.6.2.4. Visualizations

Since *All of Us* policy stipulates that we cannot report any participant-level data, we were unable to show individual trajectories without adding noise, which obscured actual trends.[141] Also, we did not include reference ranges on our mean trajectory plots due to visual overstimulation, however they were used to compile the summary in Table 10.

3.6.3. Future directions

Clusters could be better phenotyped by conducting a phenome-wide association study (PheWAS) with other fields available in *All of Us*, such as demographics (age, sex at birth, gender, race, ethnicity), comorbidities (hypertension, CVD, CKD, neuropathy, ophthalmological complications), exposures (concomitant and previous medications, substance use, physical activity), and social determinants of health (employment, housing security, food security, stress, health literacy, physical environment). Similarly, a GWAS could be utilized to characterize any genetic differences between clusters.[163] Association testing could inform feature prioritization for development of models predictive of post-treatment outcomes of interest. Also, as mentioned in Section 3.6.1.3, it may be prudent to assess the value added when cluster membership is included in predictive models compared to those based only on simple aggregate measures of longitudinal variables (quantiles, percent change).

The methods selected exclude almost every patient, introducing significant, complex selection bias. To increase cohort size, improve generalizability of results, and capture discontinuation, medication switching, and combination therapies, we could code each antidiabetic medication as its own longitudinal variable. At each time point, a participant could be assigned a value (binary,

for whether they're prescribed the medication or not, or specific dosage if available) representing their assumed state of exposure at that time.[87] This would reduce the number of participants excluded due to not having a long enough consecutive GLP-1 prescription length. We could also consider evaluating longitudinal missingness trends for each lab measure, as these patterns can be informative of underlying subpopulation differences.[164,165]

This EHR-based exploration of GLP-1 RA treatment response would be well-complemented by a similar analysis of clinical trial outcome data. Such data would likely be characterized by greater regularity of measurements and lower discontinuation rates. It may also include longitudinal patient-reported outcomes, which were not available in *All of Us*. Our analytical approach can also be extended to other treatment contexts where participants and caregivers are interested in multiple response outcomes.

3.7. Conclusion

In Aim 1, we constructed a novel research direction by combining (1) advances in deep MVTS clustering, (2) treatment response subgrouping of a medication class that is rapidly popularizing, and (3) availability of routinely collected longitudinal physical health and lab measures in the EHR. We believe this area of bioinformatics is ripe for further exploration, because as chronic disease prevalence continues to increase, novel medications with pleiotropic effects are becoming increasingly available, and temporal dynamics of medication usage are becoming easier to track and analyze via real-world data sources.

We demonstrated the ability to deploy CRLI in a complex data environment, plagued by issues like missingness, irregular time intervals, and nonstandardized measurements. However, this came at the cost of a more inclusive cohort selection strategy which could have given us greater confidence in the generalizability of our results. Requiring (1) prescription length and (2) repeat measurement minimums was necessary for building an adequately dense MVTS dataset. But it meant the final cohort may not have faithfully represented the overall GLP-1 RA medication user population. This problem may be partially alleviated over time as (1) more patients start and stay on these medications, (2) more *All of Us* data releases become available, and (3) other, complementary digital health technologies are incorporated into the *All of Us* data stream.[166]

However, while the EHR may capture the typical care patterns of chronic disease patients, EHR data is inherently plagued by the previously described issues, which can be hurdles to selecting a cohort with enough repeat measure availability to identify temporal patterns. What data sources like clinical trials, registries, and prospective observational cohorts lack in reflecting real-time, typical patient care, they make up for in high retention rates, regular measurement intervals, and detailed protocols which ensure assessment consistency and reproducibility. As such, these

sources are likely to suffer less from the specific cohort selection pressures required to build datasets amenable to analysis with CRLI and related methods.

In Aim 2, we explore one such data source, The Adolescent Brain Cognitive Development (ABCD) Study. This NIH longitudinal observational study of almost 12,000 American youths includes detailed yearly assessments encompassing hundreds of variables. Though we do not study the same biomedical context (chronic disease treatment response) in this next aim, our intent is to assess the ability of CRLI to detect meaningful multivariate trajectories in another longitudinal biomedical context that complements our findings from the EHR.

4. Aim 2: Assessing the ability of CRLI to detect meaningful trajectories in a high-dimensional, multimodal, prospective dataset (The ABCD Study)

4.1. Introduction

In our pursuit of characterizing the utility of MVTS clustering methods in longitudinal biomedical contexts, Aim 1 covered the setting of using data from electronic health records. We showed that CRLI is able to identify distinct trajectories of GLP-1 treatment response in routine clinical care. However, electronic health records, and studies based on them, can suffer from biases, including confounding, selection bias, measurement bias, and time-related bias (see Section 2.3.2 for a discussion on types of biases).[58] These limit the comparability of participant trajectories, which can vary by length, observation regularity, and missingness.

In Aim 2, we sought to assess the ability of MVTS clustering methods to detect meaningful trajectories in a longitudinal observational cohort with high dimensionality and multimodality. Our dataset of choice was the Adolescent Brain Cognitive Development (ABCD) Study, which was prospectively designed and thus does not suffer from some of the limitations of longitudinal EHR data, like irregular measurement spacing and high sparsity. The ABCD study protocol specifies a yearly follow-up schedule which most participants adhered to. While retention in the ABCD study is not perfect, an analysis of late visits, missed visits, and withdrawals by Feldstein Ewing et al. revealed no pervasive patterns of selective attrition. Out of 49,529 scheduled visits from 11,878 youth, 3.9% of visits were missed and 1.1% of participants withdrew.[167] Furthermore, the study includes assessments across an array of health-promoting factors, including environmental, genetic, neurobiological, and behavioral.[168]

Thus, ABCD is an ideal candidate for our intention to explore longitudinal observational cohorts and serves as a complement to Aim 1 in our overall characterization of MVTS clustering in

biomedical research data exploration. Aim 2 builds on the results from Aim 1 by deploying MVTS clustering methods in another important biomedical context. More specifically, prior trajectory work on ABCD has utilized group-based multi trajectory modeling, which is limited in its ability to capture nonlinear trends and requires a separate imputation strategy, which can affect subgroup identification performance. One-stage MVTS deep clustering methods are more performant and represent the most latest advancement in trajectory subtyping.[28]

Our informatics goal for this aim was to apply MVTS clustering methods to large, multimodal, longitudinal observational data. Our choice of dataset (ABCD) motivated our biomedical goal for this aim, which was to identify latent multivariate trajectory subtypes of adolescent development. We intend for our work to accelerate adolescent development-related hypothesis generation and prioritization. We posit that (1) children develop according to variable multimodal trajectories, (2) CRLI can identify these trajectories, (3) certain trajectory subgroups are enriched for negative outcomes, and (4) risk factors associated with trajectories of concern can guide surveillance and interventions.

4.2. Background and significance

We aimed to maximally utilize the longitudinal data available in ABCD to reveal common patterns of cognitive, behavioral, and physical development across several domains of adolescent health. To do this, we employed a highly performant IMVTS deep clustering approach, CRLI, capable of both imputing missing time series values and uncovering latent trajectory subgroups. We believe that identification of such patterns can inform further biomedical, clinical, and therapeutic action.

4.2.1. Multifaceted nature of adolescent development

Adolescence is a life period marked by significant multifaceted development across physical, mental, behavioral, and cognitive domains of health. Rates of mental health-related conditions, like anxiety, depression, and suicidal behavior, are on the rise worldwide. Symptoms of these conditions can emerge as early as preadolescence and often persist into adulthood if not addressed.[23,169,170] These disorders frequently co-occur and can have overlapping diagnostic criteria.[171,172] Unique symptom constellations indicative of future adverse outcomes may go undetected due to misalignment with current diagnostic paradigms. We would like to advance understanding of risk factors associated with adolescent psychiatric illness by looking at developmental trends over time.

4.2.2. Deep learning-driven clustering advances

CRLI, and other recently developed one-stage IMVTS deep clustering methods, have been successful in identifying latent longitudinal subgroups even in the context of variable missingness characteristic of high-dimensional biomedical datasets. In particular, CRLI can capture nonlinear temporal dynamics that model-based methods, like latent class growth analysis (LCGA) and group-based trajectory models (GBTM), may be unable to.[33] It operates across many variables simultaneously and handles imputation and clustering within a unified framework. The creators of CRLI, Ma et al., demonstrated its performance on gene expression data (cancer subtyping) and

medical time series (days post-surgery; ICU). Though it has not been applied to longitudinal observational cohort data, the previous state of the art one-stage method, VaDER, identified subgroups in that context, and CRLI generally outperforms it head-to-head. CRLI also outperforms other state of the art two-stage clustering methods, like k-Shape, Deep Temporal Clustering (DTC), and Deep Temporal Clustering Representation (DTCR). See Section 2.4.3.3 for further discussion on CRLI performance.

4.2.3. ABCD: longitudinal, high-dimensional, multimodal

The Adolescent Brain Cognitive Development (ABCD) Study is an ongoing, large-scale, multimodal, multidomain, longitudinal, and demographically diverse observational study which aims to track developmental patterns in almost 12,000 participants from childhood (age 9) into young adulthood (age 19). ABCD data encompasses many important developmental domains, including physical health, mental health, neurocognition, neuroimaging, and substance use. The timing of the study (2016-2027) overlaps with the COVID-19 pandemic and rapid adoption of social and electronic media among youths.[173] The sample size and data breadth and depth of ABCD are unprecedented and motivate the application of powerful analyses that can handle such high-dimensional biomedical data.[15,174]

From an informatics and biomedical data standpoint, ABCD is a valuable dataset because of its (1) high-dimensionality across risk factors, longitudinal variables, and outcomes, (2) mixture of levels of measurements (nominal, ordinal, interval, ratio), (3) many complementary modalities, including neuroimaging, biospecimens, physiological measures, clinical assessments, self- and caregiver-reported measures, and neurocognitive tests, and (4) extremely low rate of withdrawal (1.1%).[167,175] For these reasons, it represents a halfway point between carefully curated clinical trials and more messy, irregularly sampled real-world data sources.

Single-timepoint predictive models for outcomes in ABCD, such as suicidal behavior and psychopathology, fail to consider the effect that symptom trajectory may have on future outcomes. Hill et al. demonstrated that social environment and behavioral measures strongly influenced AI model-predicted psychiatric illness risk in adolescents.[176] Harman et al. found that model features important for identifying children endorsing suicidal ideation from healthy controls included feelings of loneliness and worthlessness, impulsivity, prodromal psychosis symptoms, and behavioral problems. They suggested that future analyses with longitudinal data could provide additional predictive power. [15]

Previous trajectory work on ABCD has generally (1) been restricted to univariate analysis, (2) investigated multivariate trajectories within a single domain (mental health), and/or (3) relied on traditional model-based techniques (see Section 4.3.4). Considering the drawbacks of the approaches taken by others thus far, and the suitability of ABCD for longitudinal analyses in general, we believe it would serve as an appropriate testbed to evaluate MVTS clustering methods, like CRLI. Given that adolescence is a time of significant developmental change during which a range of transitional patterns are possible, longitudinal, multidomain, multimodal data, and methods that can adequately handle that data, are crucial.

Thus, we propose to use recently developed deep learning methods (CRLI) to find groups of participants with similar patterns of longitudinal variation across developmental data domains in ABCD. These domains include physical health (anthropometrics, hormone salimetrics), mental health (Child Behavior Checklist, Brief Problem Monitor), and neurocognition (NIH toolbox). We will select instruments representative of each domain based on longitudinal availability and overall range of variance. We will iterate on cohort selection criteria and cluster validation indices to optimize CRLI performance. We will then assess associations between the identified groups and participant outcomes (psychopathology symptoms and diagnoses, drug use, risk-taking

behaviors) as indicators of adolescent mental health and behavioral challenges, a key component of overall well-being.

4.3. Related work

A number of studies have leveraged the longitudinal nature of ABCD, and other adolescent datasets, to investigate age-related changes in development across domains like physical health, mental health, and neurocognition. However, to date, one-stage IMVTS clustering methods have not been deployed in this setting. Here we discuss (1) our specific developmental domains of interest, (2) the general longitudinal structure and potential of ABCD, and (3) prior work to identify multivariate trajectories in these domains and others.

4.3.1. Self-injurious behavior during adolescence

Self-injurious behavior (SIB) is a comprehensive term, sometimes used interchangeably with suicidal thoughts and behaviors (STBs), which includes suicidal ideation (SI), suicide attempt (SA), and nonsuicidal self-injury. Several studies have characterized prevalence and predictors of overall SIB and its components, including nonsuicidal self-injury, suicide ideation, suicidal intent, and suicide attempts, in the ABCD cohort.[11–14] Harman et al. used features implicated in the adult suicide literature to build predictive models for suicidal ideation and attempt in the ABCD baseline assessment.[15] Using data from ABCD Release 4.0 (three yearly assessments), Ortin-Peralta et al. sought to characterize the transitions of suicide ideation and attempts over three years.[16] Wallace and Conner applied longitudinal network analysis (Panel Graphical Vector Autoregressive models) to evaluate associations between STBs and previously identified risk and protective factors (mental health symptoms, socioenvironment, stressors, substance use).[17]

4.3.2. The longitudinal richness of ABCD

Most major aspects of the ABCD Study design, recruitment, and dataset organization have been described extensively elsewhere.[168,177–181] We would like to highlight two papers in particular which touch on the longitudinal aspect of the dataset. In their 2021 paper, Barch et al. update their earlier 2018 report on demographic and mental health assessments in ABCD to include age-related trajectories of dimensions of psychopathology in Release 3.0 (ages 9-13). They explore trajectories of mental health (externalizing and internalizing symptoms) as a function of demographic factors, including sex, caretaker education, race and ethnicity, and socioeconomic status, using mixed effects models.[182] Hawes et al. provide an excellent discussion of many aspects of longitudinal analysis of the ABCD Study, including insights which are extendable to other rich, large-scale, longitudinal datasets. Among others, they touch on analysis of two timepoints versus three or more, typical types of stability and change in developmental trajectories, continuous and discrete outcomes, longitudinal measurement invariance, missing data and attrition, and various approaches to longitudinal analysis. They specifically note the limitations of variable-centered analyses, like those performed by Barch et al., and discuss ways in which person-centered models can address those limitations.[63] Variable-centered approaches (regression, factor analysis) focus on describing relationships between variables, whereas person-centered approaches (cluster analysis, latent class analysis) focus on relationships among individuals, aiming to classify individuals into distinct groups.[183]

4.3.3. Pubertal timing and weight status during adolescence

In their review of the role of puberty as a risk factor for STBs, Ho et al. discuss pubertal hormones (estradiol, progesterone, testosterone, dehydroepiandrosterone [DHEA]) in depth and explore how pubertal timing and sex differences might contribute to STBs in adolescence.[18] Bendezú et al. analyzed joint cortisol-DHEA trajectories in a small cohort of adolescents (n = 215, Adolescent Emotion Study) and investigated longitudinal links to emotional and behavioral problems. They

found that subgroups with similar cortisol-DHEA stress responsivity had variable associations with development of psychopathology.[19] In ABCD, Dehestani et al. investigated the effect of pubertal timing relative to chronological age, which they termed as the "puberty age gap". Their combined analysis of biological and physical pubertal features (hormones and Pubertal Developmental Scale measures) revealed an association between pubertal timing and mental health problems. This work, along with other non-ABCD studies, emphasizes the nonlinear nature of relationships between age and pubertal development.[20–22]

Herting et al. reported differences in pubertal status by sex and body weight, among other sociodemographic characteristics, in ABCD's baseline assessment (9-10 year-old children).[184] Gray et al.'s regression analyses to identify BMI predictors in the ABCD cohort found attention problems and matrix (abstract) reasoning (inversely associated); and social problems and screen time (positively associated) to be among the most important.[185] Adise et al. assessed the relationship between changes in BMI and executive function (EF) from baseline to year 2 and found that underlying differences in EF and cognition may precede weight gain.[186]

4.3.4. Multivariate trajectory modeling of adolescent development

Typically, trajectory clustering work takes the form of one of the following hypotheses:

(1) Risk factor X is associated with trajectories of variables 1, 2, …, N
(2) Trajectories of variables 1, 2,...,N are associated with outcome Y

Several groups have modeled univariate trajectories in ABCD, including those of depressive symptoms, psychotic-like experiences, addictive screen use, and cognition and neural metrics.[169,187–191] To date, Voepel-Lewis and colleagues are the only group to have modeled multivariate trajectories in ABCD. Their work has focused on co-occurring pain, psychological,

and somatic symptom trajectories, as measured by parent-reported Child Behavior Checklist (CBCL). In particular, they have explored associations of these trajectories with sociodemographics, family adversity, health care utilization, early onset substance use.[23–25] Others have applied multi-trajectory modeling to adolescent development in other datasets, exploring longitudinal trends of physical activity, mood profiles, stress experience, screen time, sleep, and more.[192–199] We have included a selection of these works in AppendixTable 1.

In sum, these works motivate further robust person-centered, multi-hormone studies that examine adolescents throughout pubertal development. Adoption of person-centered models and capture of intra-individual variation in pubertal development is critical for understanding hormonal contribution to risk for STBs in adolescents. As discussed in Chapter 2, CRLI and other one-stage methods address limitations of traditional model-based approaches that ABCD researchers have relied on, including (1) need for a strategy to deal with missing values (imputation or otherwise), (2) inability to capture complex nonlinear relationships, and (3) assumption that the form of trajectory patterns and random effects are correctly specified.[35,53]

4.4. Methods

4.4.1. ABCD Study and dataset organization

The ABCD study recruited a representative sample of 11,868 children aged 9-10 years old (together with their parents/guardians) from 22 sites across the US for biannual follow-up until age 20. Data collected include include genotype (~500k single nucleotide polymorphism variants) and more than 80,000 assessments spanning phenotype and exposures (neuroimaging, biosamples, physical exam, self-reported and parent-reported mental health measures, neurocognitive testing, and linked geospatial data). Data collection protocols, geographical distribution of research sites, and other details are available on the ABCD website.[200] The study has been described in detail elsewhere.[179,181,201]

We used data from ABCD Release 5.0 (June 2023), in which events (also referred to as timepoints) up to the 3-year follow-up event were completed, with varying numbers of missed visits per event.[202,203] The 42-month and 4-year follow-up events were still ongoing when the data for this release was frozen, so data are included for a subset of participants for these timepoints.[204] In total, release 5.0 contains data for 9 timepoints: 5 annual in-person visits and 4 mid-year remote check-ins. Figure 39 shows the age distribution and participant count at each event.

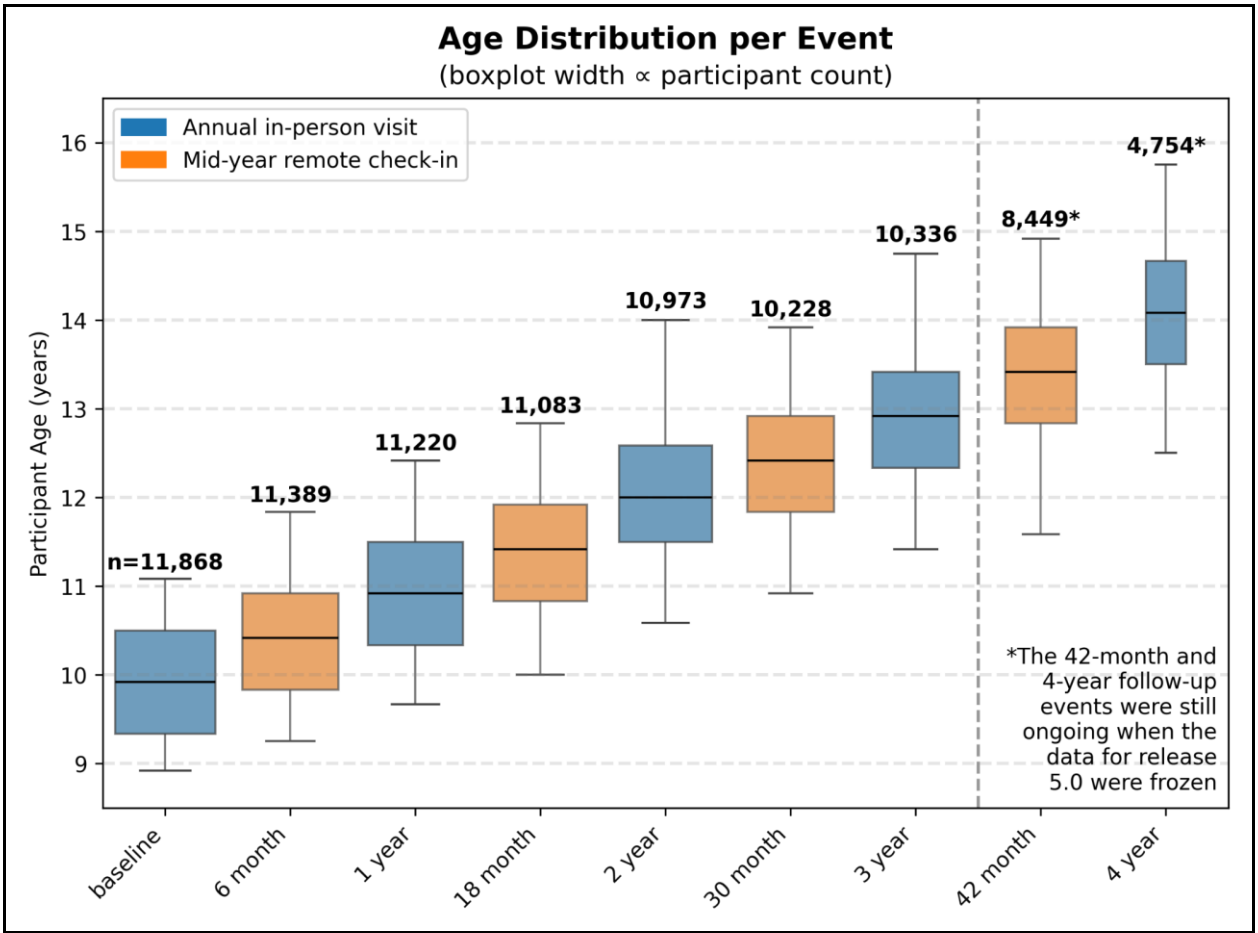


Figure 39. Participant count and age distribution at each ABCD assessment event

Table 11 shows that assessment measurement and data collection vary between ABCD events. Importantly, mid-year events were conducted on the phone or online and were significantly less comprehensive than annual in-person events. Thus, the vast majority of variables were only collected at a subset of annual timepoints.

	Baseline (in person)	6 months (phone / on-line)	**Year 1 (in person)**	18 months (phone / on-line)	**Year 2 (in person)**	30 months (phone / on-line)	**Year 3 (in person)**	42 months (phone / on-line)	**Year 4 (in person)**	Ongoing ….
Participant Age	9-10 yrs		10-11 yrs		11-12 yrs		12-13 yrs		13-14 yrs	
Substance Use & Related Factors										
Youth Self-Report on Substance Exposure/Use	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Related Risk and Protective Factors	✓		✓		✓		✓		✓	
Saliva & Hair Samples	✓		✓		✓		✓		✓	
Mental Health & Related Factors										
Parent-Report on Youth	✓		✓		✓		✓		✓	
Youth Self-Report	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Parent Self-Report	✓									
Parent Report on Family	✓									
Physical Health & Related Factors										
Parent-Report on Youth	✓		✓		✓		✓		✓	
Youth Self-Report	✓		✓		✓		✓		✓	
FitBit					✓				✓	
Saliva for Hormone Assessment	✓		✓		✓		✓		✓	
Saliva/Blood for DNA & Health Factors	✓				✓				✓	
Neurocognition										
NIH Toolbox	✓				✓				✓	
Other “cold” cognitive measures	✓				✓		✓		✓	
Other “hot” cognitive measures			✓				✓			
Gender Identity and Sexual Health	✓		✓		✓		✓		✓	
Culture and Environment										
Parent-Report on Youth/Family	✓		✓		✓		✓		✓	
Youth Self-Report	✓		✓		✓		✓		✓	
Geocoding / Neighborhood Factors	✓		✓		✓		✓		✓	
Brain Imaging	✓				✓				✓	

Note. ✓ = Questionnaires and interviews; ✓ = biospecimen samples (e.g., saliva, hair, blood); ✓ = Fitbit; ✓ = saliva sample for hormone assessment; ✓ = NIH Toobox; ✓ = “cold” cognitive measures; ✓ = "hot" cognitive measures; ✓ = geocoded factors; ✓ = brain imaging measures.

Table 11. Data collection differences between assessment timepoints[201]

Data in ABCD are organized in a hierarchical structure by assessment domain, then table, then individual assessments (also referred to as variables). For example, the Mental Health (MH) assessment domain contains table *mh_p_cbcl* which includes all 201 parent-reported Child Behavior Checklist (CBCL) assessments. These assessments could be single instruments, unidimensional composites, or multidimensional composites (Figure 40). In the case of CBCL, a single instrument would be one of the 119 behaviors (e.g., “Feels they have to be perfect”) presented to parents/caregivers to be rated 0, 1, or 2 (not true, sometimes true, often true) depending on the frequency with which their child exhibited that behavior over the preceding 6 months. These individual behavioral assessments can be summed to create eight unidimensional

composite syndrome scales (anxious/depressed, withdrawn/depressed, somatic complaints, etc). These can be further combined to create multidimensional composite syndrome scales; anxious/depressed, withdrawn/ depressed, and somatic complaints scales are rolled up to create a broader internalizing problems composite (Table 12).[205–207] This paradigm can be extended to other assessment domains (e.g., domain: physical health, single instruments: weight & height, composite: BMI).

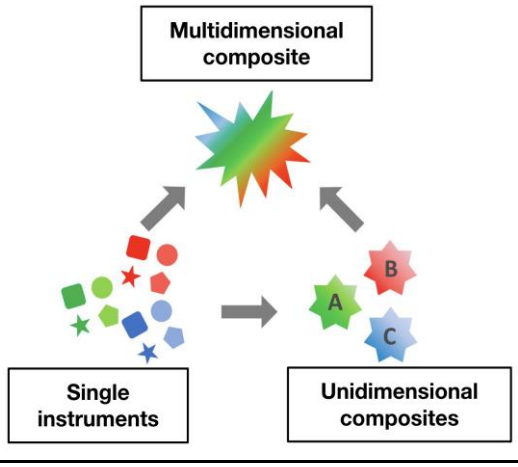


Figure 40. From single instruments to multidimensional composites[106]

Total Problems							
Broad-Spectrum Groupings							
Internalizing			Neither Internalizing nor Externalizing			Externalizing	
Syndromes							
Anxious/ Depressed	Withdrawn/ Depressed	Somatic Complaints	Social Problems	Thought Problems	Attention Problems	Rule-Breaking Behavior	Aggressive Behavior
Examples of Problem Items[a]							
Cries	Enjoys little	Aches, pains	Clumsy	Can't get mind off thoughts	Acts young	Bad companions	Argues
Fears	Lacks energy	Eye problems	Gets teased		Can't concentrate	Breaks rules	Attacks people
Feels unloved	Rather be alone	Feels dizzy	Jealous	Hears things	Can't sit still	Lacks guilt	Disobedient
Feels too guilty	Refuses to talk	Headaches	Lonely	Repeats acts	Confused	Lies, cheats	Fights
Talks or thinks of suicide	Sad	Nausea	Not liked	Sees things	Daydreams	Sets fires	Loud
Worries	Secretive	Overtired	Prefers younger kids	Strange behavior	Fails to finish	Steals	Mean
	Shy, timid	Stomach-aches	Too dependent	Strange ideas	Impulsive	Swearing	Screams
	Withdrawn	Tired			Inattentive	Truant	Temper
		Vomiting			Stares	Vandalism	Threatens

[a]Abbreviated versions of CBCL/6-18, TRF, and YSR items.

Table 12. Hierarchical model of empirically based assessment levels[207]

4.4.2. Longitudinal feature selection

ABCD 5.0 contains 86,591 variables with irregular longitudinality. Three examples of longitudinal variables and their participant counts are shown in Figure 41. We systematically assessed the

longitudinal nature of every non-imaging variable in ABCD by (1) counting how many events they were measured at and (2) for how many study participants they were measured (AppendixTable 2). The only instruments measured at more than 5 timepoints were the NIH Toolbox Positive Affect (POA) and the Brief Problem Monitor (BPM).[182,208,209]

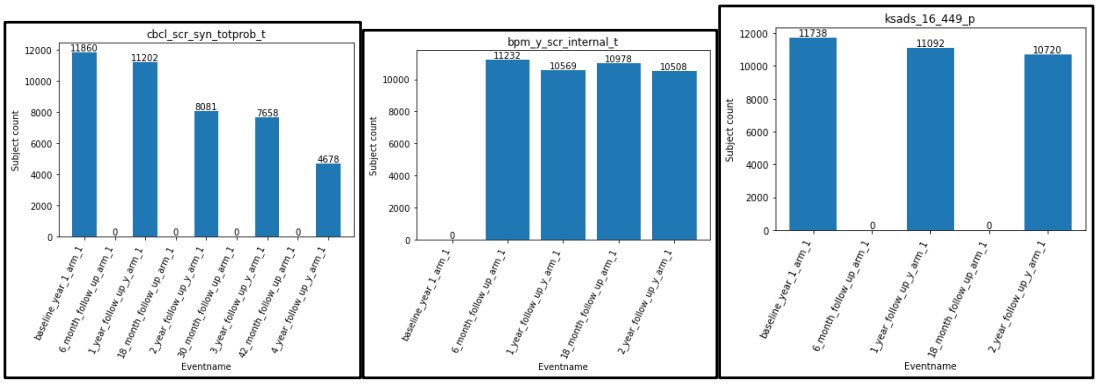


Figure 41. Representative examples of irregular longitudinality between variables

A basic feature selection question arises when presented with hundreds of longitudinal variables to select from: which will be most likely to inform meaningful multivariate trajectories? We considered several philosophies that could drive our variable selection: prior knowledge, clinical value, similarity, and data characteristics. Prior knowledge-driven selection relies on already published literature pointing to certain variables that have interesting longitudinal patterns in adolescents. Though this philosophy would allow us to use MVTS clustering to validate past findings, it would not generate novel insights or generate new hypotheses. Clinical value-driven selection prioritizes variables that are most clinically meaningful for pediatricians, child psychologists, suicidologists, etc. This selection would be informed by our clinical collaborators based on their own knowledge of the literature and their personal experiences working with children and adolescents. Similarity-driven selection builds on intrinsic variable similarities and potential biological or pharmacological insights that studying these variables together could entail. For example, studying variables within the same domain (mental health, physical health, neurocognition, etc.) or the same modality (salivary biosamples, anthropometrics, CBCL scales)

is likely to be more biologically and clinically interpretable than cross-domain or cross-modality analyses. Lastly, data characteristics-driven selection is a prior knowledge-agnostic approach that prioritizes variables based on longitudinal availability and potential for statistical variation over time. What this approach lacks in interpretability it makes up for in ability to generate novel insights about multidomain, multimodality developmental patterns. This approach is also the most reproducible as it relies only on innate properties of the data rather than external expertise or notions of value. These philosophies are not necessarily mutually exclusive; we pursued a combination of these four philosophies, with methodological emphasis placed on variability-driven and domain-based selection.

4.4.2.1. Variability-driven selection

We first attempted to take a purely data-driven approach by prioritizing variables which had high coefficients of variation (CV) across timepoints. CV, also known as the relative standard deviation, is defined as the ratio of the standard deviation to the mean. It is a relative measure of variability that allows for comparison between variables with different units, means, and distributions. For each participant, we calculated the CV for each quantitative variable across all timepoints. Then we generated ordered boxplots to visualize the distribution of CVs for each variable (AppendixFigure 1). In financial time series analysis, CV is a metric used to assess volatility. We anticipated that a variable with higher mean CV across participants would have more interesting patterns of variability from which MVTS clustering methods could glean unique trajectory patterns.

AppendixTable 3 summarizes the four basic levels of measurement for a given variable. A variable's level reflects its intrinsic statistical properties, possible measures of central tendency and variability, and most appropriate visualization. Of note, at time of writing, the ABCD Study does not have a recommended strategy to distinguish numerical (quantitative) from categorical (qualitative) variables (though this has since been addressed in Release 6.0, see Section

6.2.3).[210] This is an important consideration because some variables may “appear” numerical without context. These variables, though recorded numerically, are purely categorical with the numbers not representing a scale of any kind. To address this, we assumed a variable was categorical if it had 10 or fewer unique values present in the dataset.

4.4.2.2. Domain-based selection

We also opted to select longitudinal features in a more “traditional” manner, drawing from ABCD’s inherent domain-centric structure, collaborator expertise, and published work.

4.4.2.2.1. Mental Health

The Child Behavior Checklist (CBCL) is an inventory of the Achenbach System of Empirically Based Assessment (ASEBA) School-Age Forms and Profiles. This parent-reported instrument assesses youth competencies and psychopathology in dimensional terms. The Empirically Based Syndromes scales were obtained through factor analysis, and items were grouped into the following eight dimensions: Anxious-Depressed, Withdrawn/Depressed, Somatic Complaints, Social Problems, Thought Problems, Attention Problems, Rule-Breaking Behavior, and Aggressive Behavior (Table 12).[207,211] Per Barch et al., we used raw scores, rather than age and sex-adjusted T scores, to better capture relationships to developmental and sex differences (Table 13).[182] These eight CBCL syndrome scales were used, as opposed to other available scales, due to their empirically derived factor structure.[212]

table_name	var_name	var_label
mh_p_cbcl	cbcl_scr_syn_anxdep_r	AnxDep CBCL Syndrome Scale (raw score)
mh_p_cbcl	cbcl_scr_syn_withdep_r	WithDep CBCL Syndrome Scale (raw score)
mh_p_cbcl	cbcl_scr_syn_somatic_r	Somatic CBCL Syndrome Scale (raw score)
mh_p_cbcl	cbcl_scr_syn_social_r	Social CBCL Syndrome Scale (raw score)
mh_p_cbcl	cbcl_scr_syn_thought_r	Thought CBCL Syndrome Scale (raw score)
mh_p_cbcl	cbcl_scr_syn_attention_r	Attention CBCL Syndrome Scale (raw score)

mh_p_cbcl	cbcl_scr_syn_rulebreak_r	RuleBreak CBCL Syndrome Scale (raw score)
mh_p_cbcl	cbcl_scr_syn_aggressive_r	Aggressive CBCL Syndrome Scale (raw score)

Table 13. ABCD mental health variable names and labels

4.4.2.2.2. Physical Health

ABCD includes measures of three pubertal hormone levels (estradiol in females only, testosterone and DHEA in all participants) from saliva.[213] These were assessed at all 5 annual timepoints. Anthropometrics (height, weight, waist circumference) were also measured annually. The specific physical health variables we extracted from the ABCD dataset are summarized in Table 14. We engineered a BMI feature by the following equation, $BMI = 703 \times \frac{weight\ (lbs)}{height^2\ (inches)}$. Of note, a total of 26 participants' height and weight measures in the 5.0 release may have been affected by data errors (decimal place shifts, transpositions, typos)[j], which we did not account for (see Section 4.6.2.2 for further discussion).

table_name	**var_name**	**var_label**
ph_y_anthro	anthroheightcalc	Standing Height Average (inches)
ph_y_anthro	anthroweightcalc	Average Measured Weight (lbs)
ph_y_sal_horm	hormone_scr_dhea_mean	Salimetrics hormone test DHEA mean of measures (pg/mL)
ph_y_sal_horm	hormone_scr_hse_mean	Salimetrics hormone test estradiol (HSE) mean of measures (pg/mL)
ph_y_sal_horm	hormone_scr_ert_mean	Salimetrics hormone test testosterone (ERT) mean of measures (pg/mL)

Table 14. ABCD physical health variable names and labels

4.4.2.2.3. Neurocognition

The NIH Toolbox for the Assessment of Neurological and Behavioral Function (NIH-TB) is a computerized, standardized assessment which measures performance across four domains: cognition, motor function, sensation, and emotion. Specifically, the NIH-TB Cognition Battery

[j] See "ABCD 5.0 Changes and Known Issues - update June 22 2023" file on the NIMH Data Archive website for more details[214]

includes seven tests designed to tap constructs within specific cognition subdomains. The tests and corresponding cognitive constructs are summarized in Table 15 and described in more detail in AppendixTable 4.[215,216] Of these seven, five were measured in more than 1,000 participants at 3 or more timepoints (starred in table).

Cognitive construct	NIH-TB test
Executive/Attention	Flanker Inhibitory Control and Attention*
Executive/ Shifting	Dimensional Change Card Sort
Working Memory	List Sorting Working Memory
Episodic Memory	Picture Sequence Memory*
Language	Oral Reading Recognition*
Language	Picture Vocabulary*
Processing Speed	Pattern Comparison Processing Speed*

Table 15. Cognitive constructs and NIH Toolbox tests

4.4.3. Outcome measures

Kiddie Schedule for Affective Disorders and Schizophrenia for School-Aged Children (KSADS) is a diagnostic semi-structured interview used to measure psychiatric disorders in children aged 6-18 years old. It assesses current and past symptoms of depression, mood, anxiety, ADHD, psychotic, and disruptive behavior disorders, among others.[217,218] Starting with the year 3 follow-up, the ABCD study switched from KSADS 1.0 to 2.0, which includes better assessments of autism spectrum and psychotic disorders. KSADS-COMP refers to the self-administered, computerized version, which was used for information gathering in ABCD.[182]

For outcomes related to broad psychopathology, we used measures from the KSADS Symptoms & Diagnoses table (*mh_y_ksads_ss*), which captures DSM-V based symptoms and diagnoses based on youth-reported responses to individual KSADS questions.[217] Children answered questions about present (past two weeks) and past (lifetime) mental health experiences. All diagnoses and symptoms were ultimately coded as 0 (absent) or 1 (present).[182]

From this table, we selected all present diagnoses and symptoms at years 3 and 4, amounting to 96 and 226 variables respectively. This discrepancy is because, according to the ABCD Study's

prespecified assessment protocol, at year 3, only KSADS modules for suicide and substance use-related behavior were collected, whereas a full KSADS was collected in year 4 (AppendixTable 5).[182,219–221]

4.4.4. Cohort selection

4.4.4.1. Longitudinal data considerations

Cohort selection was driven primarily by availability of repeat measures of variables of interest. This availability was affected by (1) study protocol, (2) documented data release issues (3) participant withdrawals, and (4) missed visits. Between data releases, instruments are re-evaluated and can be found to be incorrect or misleading. For example, certain KSADS diagnostic variables were not included in data release 4.0 because a review conducted by KSADS-COMP originators revealed errors in diagnostic criteria which likely let to overestimates of several disorder diagnoses.[182,214] The ABCD Retention Workgroup has characterized missing visits, late visits, and participant withdrawals from the study up to and including year 3 follow-up.[167] Retention patterns affect the availability of longitudinal variables suitable for MVTS clustering.

4.4.4.2. Which subset of timepoints?

In the context of prediction models, trajectory cluster labels can serve as model features and/or model outcomes. For example, we could hypothesize that exposure X at baseline predicts trajectory membership across the next 3 years. On the other hand, we could hypothesize that inclusion of trajectory cluster label (y0-3) in a prediction model of some y4 outcome, like SIB, will improve model performance compared to a single-timepoint model. To that end, we considered three subsets of timepoints when selecting our cohorts:

(1) y0-y3: to find associations between pre-year 4 trajectories and year 4 outcomes (since year 4 was the last available data release)

(3) y0-y2: to find associations between pre-year 3 trajectories and year 3 outcomes (since year 3 was the last full data release)

(2) y0-y4: to find associations between full trajectories and year 4 outcome (since this approach made full use of all available longitudinal data)

4.4.4.3. Sex differences

In clustering experiments involving physical measures (pubertal hormones, anthropometrics) we clustered males and females (sex at birth) separately due to (1) estradiol being measured only in females in the ABCD study and (2) documented differences in pubertal development between sexes.[184] Sex differences in mental health and behavioral patterns and outcomes have also been well described in previous studies.[182,222] The ABCD variable "demo_sex_v2" captures sex at birth by asking "What sex was the child assigned at birth, on the original birth certificate?". Sex assigned at birth differs from gender identity, though we did not investigate this relationship. Previous work has applied the 2-step method to ascertain that 238 (5%) of youths identify as gender diverse at year 4.[223]

4.4.4.4. Self-injurious behavior

Working backwards, we aimed to choose an outcome representative of adolescent well-being, like self-injurious behavior (SIB), and identify multivariate trajectories within the group of participants that exhibited that behavior at year 3. Using the KSADS diagnoses table mentioned in Section 4.4.3, we developed an aggregate binary SIB measure based on all ten available youth-reported, present suicidality diagnoses at year 3 (participant counts for each diagnosis are shown in Table 16).[224] If a participant had any of these ten measures present at year 3, they were included in the SIB cohort (n = 441).

Absent	Present	Missing	var_name	var_label
9757	441	0	suicidal_agg	Aggregate across ALL suicidality diagnoses (below)
9941	257	0	ksads2_23_905_t	SelfInjuriousBehaviorwithoutsuicidalintentPresent
9947	251	0	ksads2_23_907_t	SuicidalideationActivenonspecificPresent
10071	127	0	ksads2_23_906_t	SuicidalideationPassivePresent
10090	108	0	ksads2_23_908_t	SuicidalideationActivemethodPresent
10140	58	0	ksads2_23_909_t	SuicidalideationActiveintentPresent
10150	48	0	ksads2_23_910_t	SuicidalideationActiveplanPresent
10165	33	0	ksads2_23_911_t	PreparatoryActionstowardimminentSuicidalbehaviorPresent
10181	15	2	ksads2_23_914_t	SuicideAttemptPresent
10187	11	0	ksads2_23_913_t	AbortedAttemptPresent
10194	4	0	ksads2_23_912_t	InterruptedAttemptPresent

Table 16. KSADS SIB aggregate binary measure at year 3

4.4.5. Data preprocessing

4.4.5.1. Ordering participants by calendar age

During adolescence, even small calendar age differences can mean drastically different stages of physical and mental development.[20] Other groups performing trajectory clustering on the ABCD dataset have used assessment periods as their timepoints of choice.[23,25] Since participant ages are not uniform at each assessment period (Figure 39), we instead opted to use calendar age. It is also a more granular, interpretable, and clinically meaningful way of ordering participants than by assessment period.[63] Also, the naming convention of the assessment periods (1 year followup, 2 year followup, etc) implies that participants have aged ~1 year between assessments, but this is not always the case (Figure 43).

Another consideration is how to group half ages. For example, should a 13.5 year old be considered 13 (as they would if they walked into a doctor's office) or 14 (as they by decimal rounding convention)? Another option is treating each half-age as its own timepoint to increase

time series length and granularity, however decreasing the number of participants at each timepoint.

Given these considerations, we tried the following approaches: (1) annual event, (2) typical age rounding, (3) age integer floor (clinical approach), (4) half years for maximum granularity. Representative examples of these approaches are shown in Figure 42.

For our final models, we used the third approach, whereby a participant with age x such that $11 \leq x < 12$ would be considered 11.

After this age rounding, in cases where a participant was the same age at consecutive timepoints, we averaged values across those timepoints.

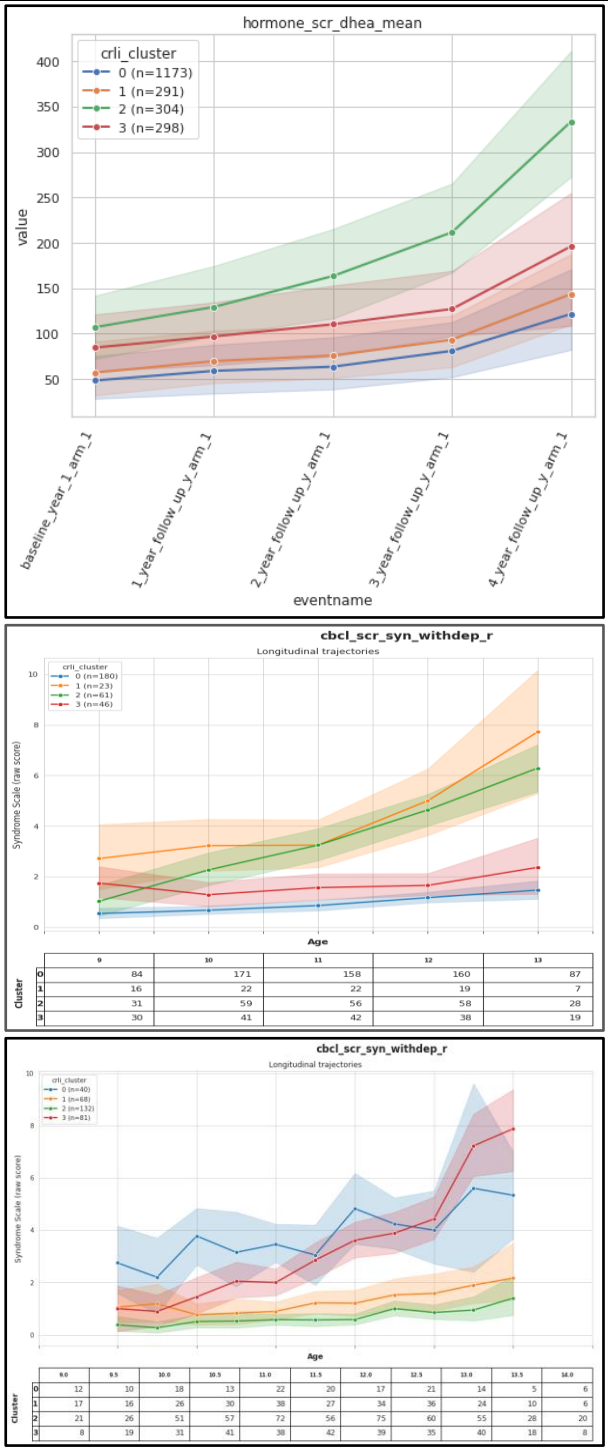


Figure 42. Different age approaches

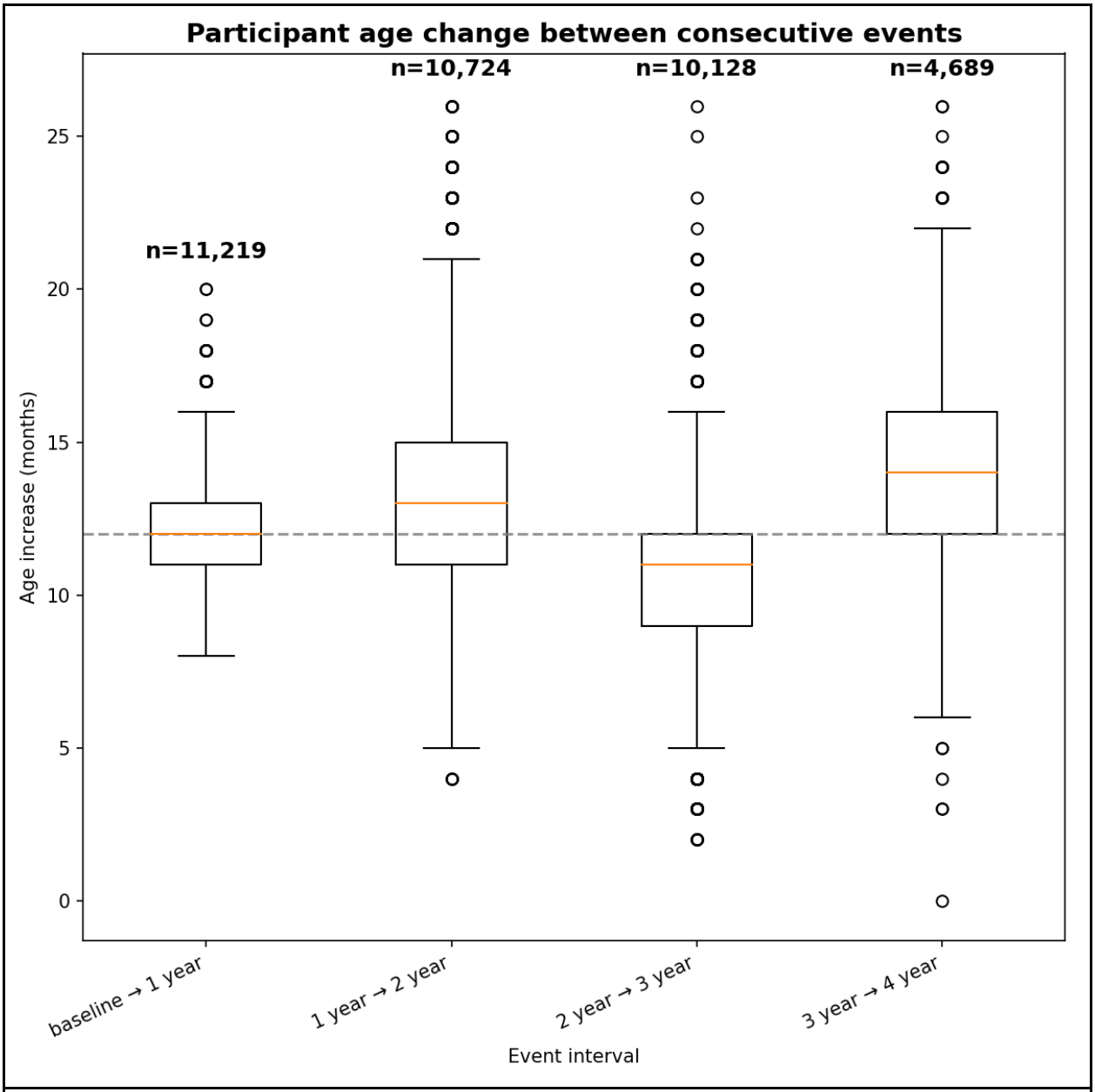


Figure 43. Participant age change (in months) between consecutive yearly ABCD assessment events

4.4.5.2. Normalization relative to baseline

Following de Jong et al., for all variables, all time points were normalized relative to baseline by (i) subtracting the baseline mean across all participants and (ii) dividing by the baseline standard deviation across all participants.[54] We calculated and saved the baseline mean and standard deviation before converting to calendar age and applied the normalization with these saved values after the conversion.

4.4.6. Clustering Representation Learning on Incomplete time-series data (CRLI)

As discussed in Chapter 3 Methods, CRLI is among the recently developed deep learning methods which simultaneously perform missing value imputation and clustering across MVTS (Figure 30). We used the PyPOTS implementation of CRLI. Due to computational constraints, we used the same default hyperparameters as in the original CRLI publication. See Section 3.4.4. for further details.

4.4.7. Analysis

4.4.7.1. Clustering validation indices

As discussed in Chapter 3 Methods, we utilized internal clustering validation indices (CVI) to select optimal cluster number since we did not have ground truth subgroups to compare against. Specifically, we aimed to maximize Calinski-Harabasz and Silhouette indices and minimize Davies-Bouldin and S_Dbw indices (Table 5).[68,84,87,89,148] See Section 3.4.5 for further details.

4.4.7.2. Outcome measures

We used the TableOne package to perform chi-squared tests to determine if outcome proportions at years 3 and 4 differed between clusters.[225] All outcomes (KSADS diagnoses and symptoms) were categorical and presented in the dataset as 0 (absent) or 1 (present).[182]

4.4.8. Visualization

4.4.8.1. Trajectories

Trajectories were visualized using (1) boxplots and (2) mean lineplots. Boxplots are more informative, complete visualizations as they show the full spread of datapoints for each cluster at each timepoint. Mean trajectories (with confidence intervals) sacrifice data completeness but are more digestible for the viewer. We added a table at the bottom of both visualization styles which includes counts of total measures taken at each timepoint. This captures the missingness in the

cohort which is not immediately evident from the cluster participant counts in the legend. Finally, we included barplots showing the average timepoint medians as a way of simplifying the cluster comparisons for each variable. However, these barplots do not capture important trajectory features like rates of change between timepoints and overall patterns of fluctuation.

4.4.8.2. Outcome measure proportions

We used stacked barplots to visualize differences in outcome proportions between clusters. Barplots are ordered by ascending unadjusted chi-squared test p-value though we do report both unadjusted and Bonferroni-adjusted p-values. TableOne internally applies the Bonferroni correction by multiplying unadjusted p-values by the total number of tests performed.[226] Barplot column widths were scaled to the number of participants the outcome was measured in. Accordingly, some outcomes are ranked highly due to low p-value but have very few observations. In Results, we show the top 9 most significant outcomes (up to 0.1 unadjusted p-value).

4.4.9. Coding workflow

Our work was limited to the "core" directory in the ABCD 5.0 release, which we downloaded from the NIMH Data Archive in 2023.[202,214] Our full analytical workflow is as follows:

Initial measure filtering:

1) Read in tables/measures of interest informed by feature selection strategy [see Section 4.4.2]
2) Engineer any new measures of interest (ex: BMI from height and weight)
3) *Optionally, drop any individual timepoints entirely (ex: year 4 due to incomplete data release)*
4) *Optionally, subset to single sex and/or specific KSADS outcome (ex: SIB present at year 4) [see Section 4.4.4]*

5) Very extreme outliers (z-scores > 10) were identified and treated as missing (replaced with np.nan)
6) Ensure all measures are recorded for some minimum number of participants (ex: 100, 500, 1,000) at 3 or more timepoints
7) Keep only those participants who have all measures recorded at 3 or more timepoints
8) Drop any rows (single participant-timepoint data) which are NA (missing) for all measures [this removes any empty timepoints for each participant but does not affect total subject count]

Age conversion and normalization:

9) Calculate baseline mean and standard deviation for each measure
10) Convert assessment timepoint to age at assessment timepoint [see Section 4.4.5.1]
11) Average measurements in cases where participant has duplicate ages
12) Drop any ages for which there are very few participants (ex: 8, 14)
13) Keep only those participants who have all measures recorded at 3 or more ages (repeat of step 7)
14) Normalize all timepoints (ages) relative to baseline by (i) subtracting baseline mean and (ii) dividing by standard deviation (recorded in step 9)[54]
    a) Save unnormalized data for plotting in step 18

Clustering:

15) Convert normalized dataframe to PyPOTS format (3D array of the shape participants*timepoints*measures)
16) For each cluster number [2,9]:
    a) Run CRLI on formatted array with default hyperparameters[227] [see Section 4.4.6]
    b) Calculate internal CVI

c) Store clustering results (predict on same data CRLI was trained on)

17) Plot internal CVI values for each cluster number to inform best cluster number selection

18) For best cluster number, generate trajectory visualizations for each measure [see Section 4.4.8.1]

Outcome measure associations

19) For years 3 and 4, separately:

a) Retrieve present KSADS symptoms and diagnoses [see Section 4.4.3]

b) Merge outcome measures from (a) with cluster assignments from step 18

c) Perform chi-square tests of independence for all outcome measures and apply Bonferroni adjustment to resulting p-values [225,228–230]

d) Visualize most significant outcome measures as stacked barplots

4.5. Results

4.5.1. Experiment #1: Trajectories of pubertal hormones and BMI in female participants

Clustering trajectories of the three pubertal hormones (DHEA, testosterone, estradiol) and BMI from ages 9-14 of female participants (n = 2,923) revealed 3 distinct subgroups.

4.5.1.1. Cohort selection

Estradiol was only measured in female participants, so we limited our cohort to females only to maximize generalizability of our results.[184] We further selected based on availability of all 4 physical health measures at 3 or more timepoints, per discussion in Hawes et al.[63] After additional preprocessing and filtering, our final cohort consisted of 2,923 participants (Figure 44).

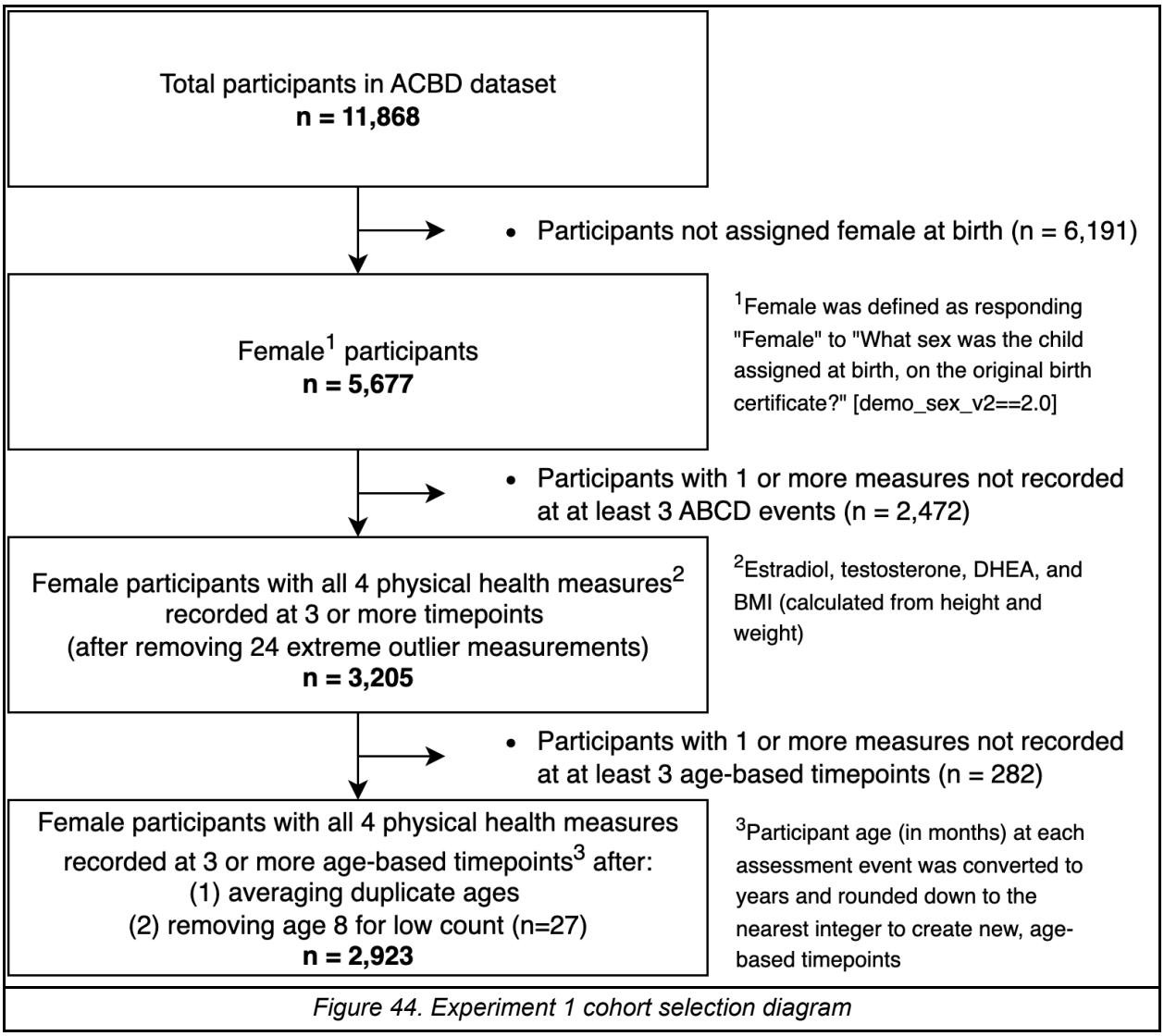


Figure 44. Experiment 1 cohort selection diagram

4.5.1.2. Optimal cluster number selection

We concluded that 3 clusters was optimal based on (i) maximization of the Calinski-Harabasz score and (ii) relative minimization of the Davies-Bouldin score and (iii) S_Dbw index. Figure 45 shows calculations of each index for CRLI results specified on 2 to 6 clusters. Red points indicate the cluster number for which the index achieved the best performance.

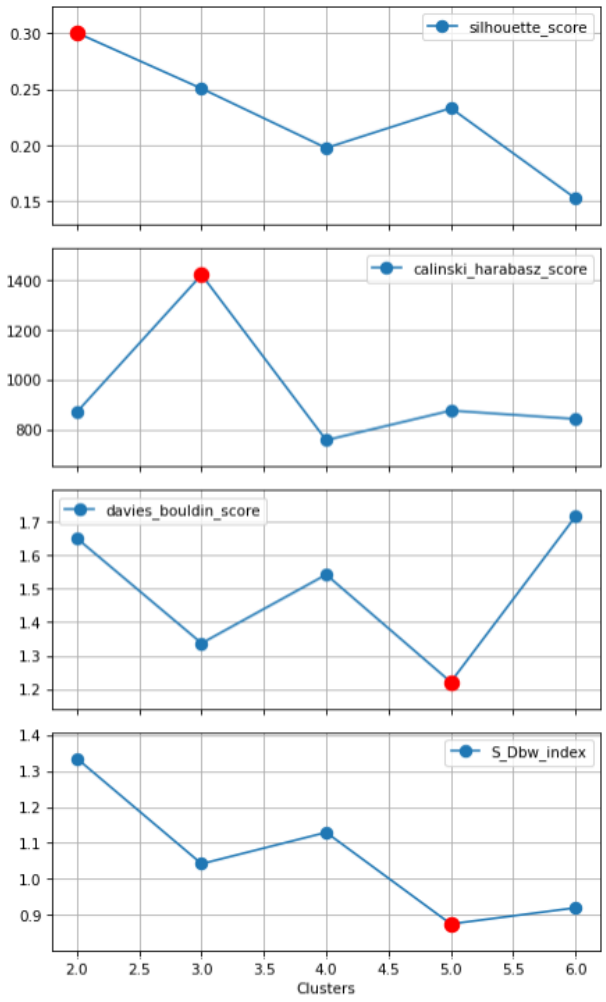
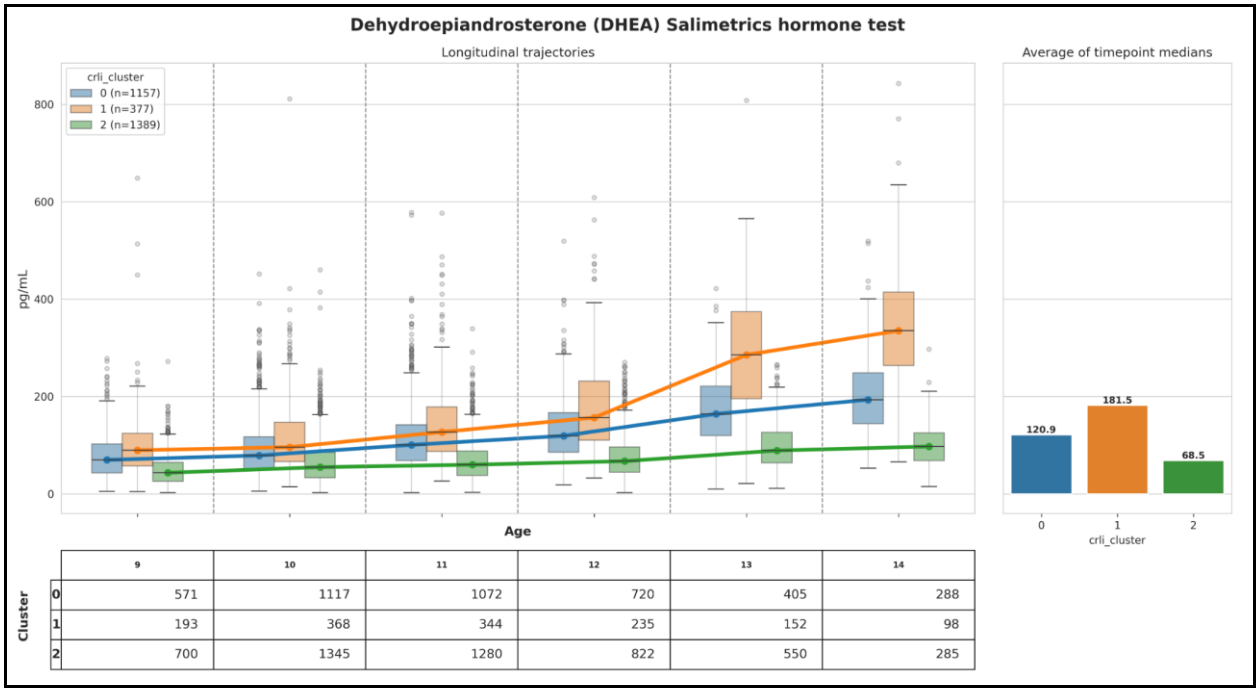


Figure 45. Internal CVI for female physical health clusters

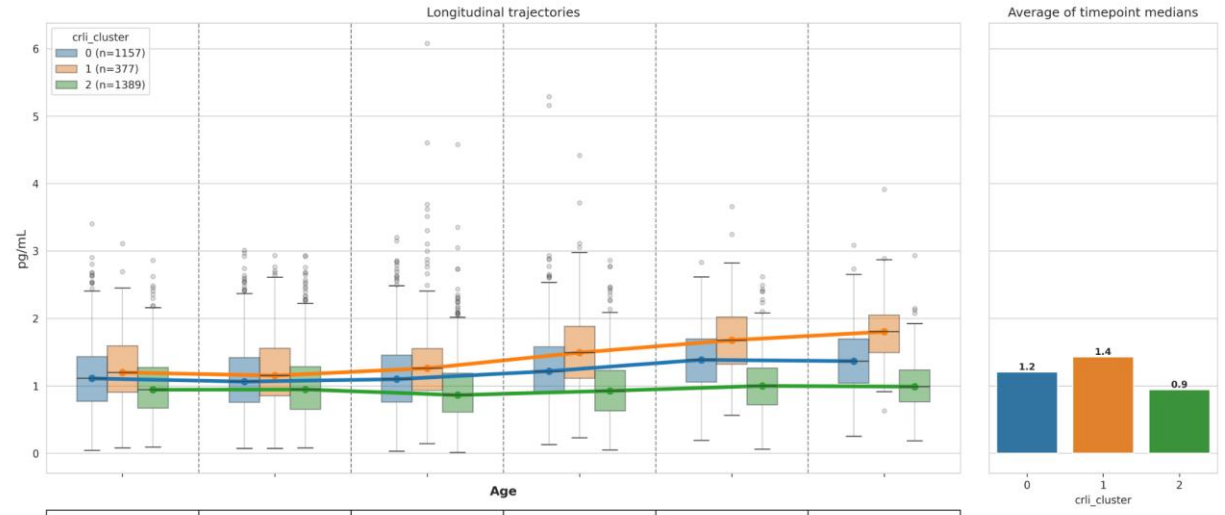
4.5.1.3. Trajectory visualization

Cluster 1 (orange, n = 377) was characterized by sharp increases in DHEA and testosterone (ERT) after age 12 and an overall higher BMI than the other clusters. Cluster 0 (blue, n = 1,157) had more modest increases in DHEA and ERT after age 12. Finally, Cluster 2 (green, n = 1,389) showed the lowest levels across all measures and overall longitudinal stability (Figure 46). Mean lineplot overlays with 95% confidence intervals are shown in AppendixFigure 2.



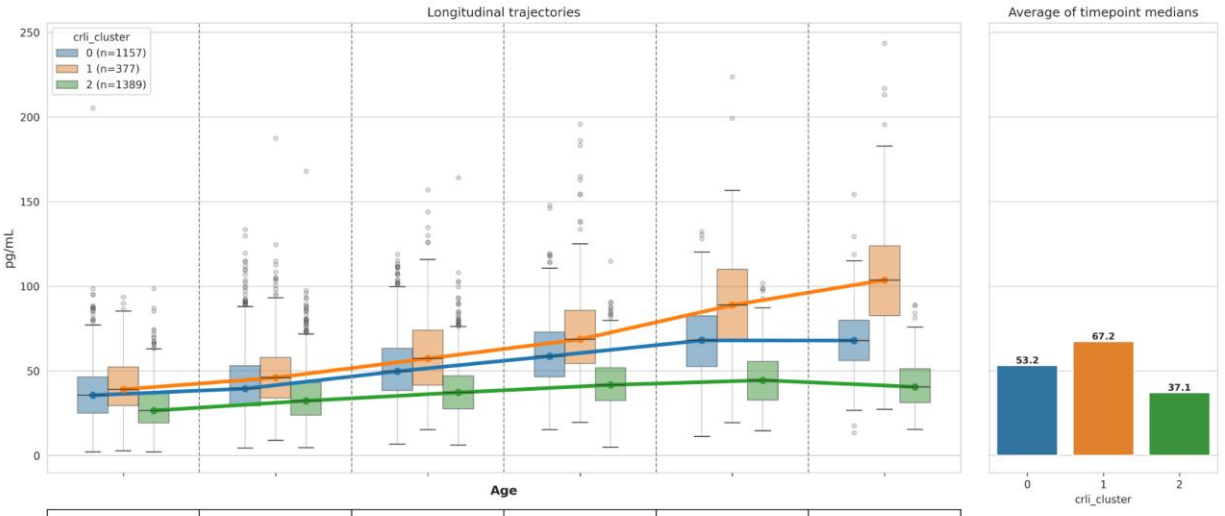
Cluster	9	10	11	12	13	14
0	571	1117	1072	720	405	288
1	193	368	344	235	152	98
2	700	1345	1280	822	550	285

Estradiol (HSE) Salimetrics hormone test



Cluster	9	10	11	12	13	14
0	571	1117	1072	720	405	288
1	193	368	344	235	152	98
2	700	1345	1280	822	550	285

Testosterone (ERT) Salimetrics hormone test



Cluster	9	10	11	12	13	14
0	571	1117	1072	720	405	288
1	193	368	344	235	152	98
2	700	1345	1280	822	550	285

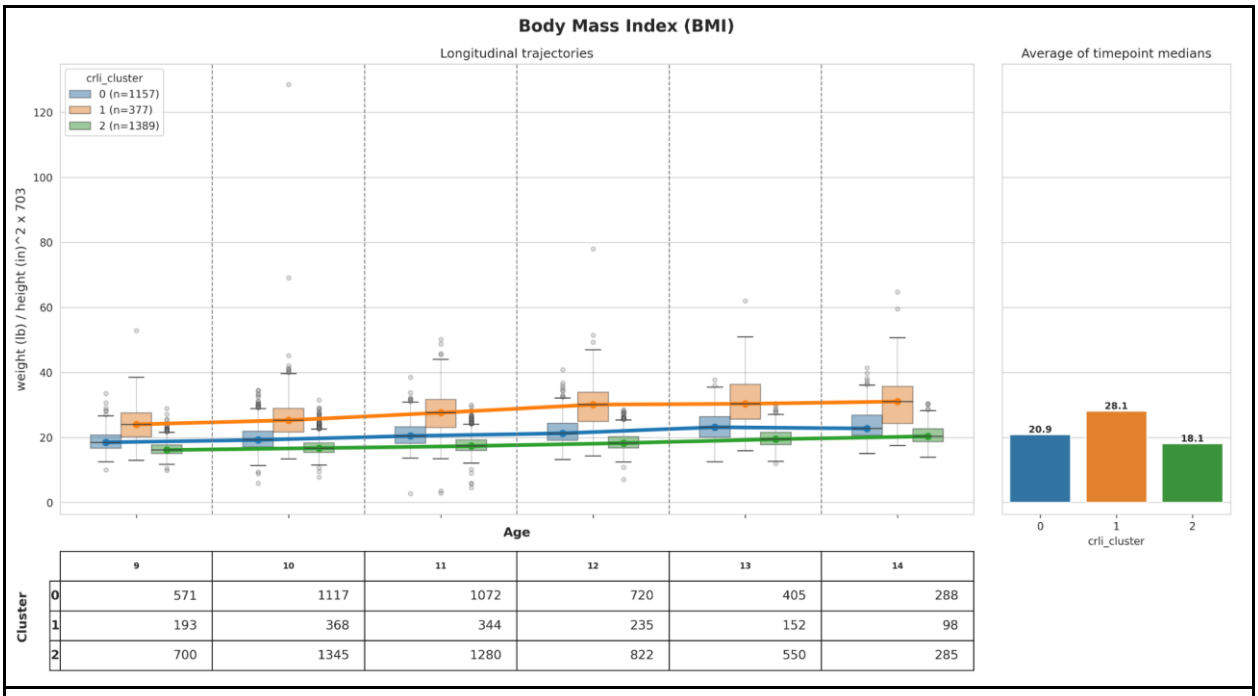


Figure 46. Multivariate physical health trajectories of female participants

4.5.1.4. Associations of trajectories with outcome measures

Top significant associations between cluster membership and Year 3 KSADS outcomes revolved mostly around suicidality and self-injurious behavior (Figure 47). Cluster 2 (green) had the lowest proportions of participants exhibiting such behaviors and Cluster 0 (blue) had the highest. Suicidal ideation and self injurious behavior were exhibited by 2.6% and 3.1%, respectively, of participants in Cluster 2. In Cluster 0, those proportions were 4.5% and 5.1%.

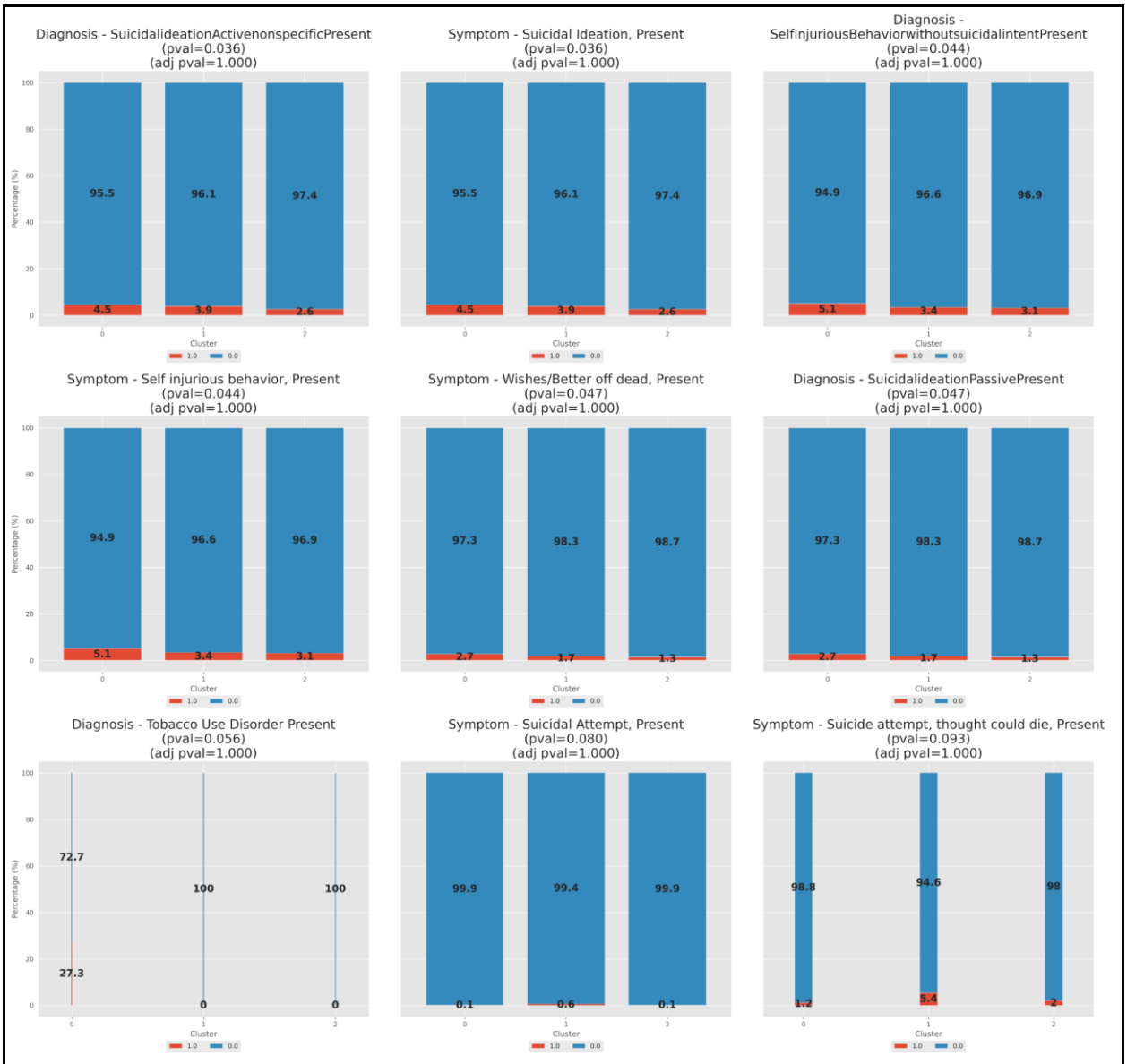


Figure 47. Year 3 outcome measure proportions per cluster[k]

Top significant associations between cluster membership and Year 4 KSADS outcomes revolved mostly around weight and eating-related behaviors (Figure 48). Cluster 1 (orange) had higher proportions of participants exhibiting fear of becoming obese, weight control vomiting, and binge eating.

[k] We report unadjusted and Bonferroni-adjusted chi-squared test p-values. TableOne package internally applies the Bonferroni correction by multiplying unadjusted p-values by the total number of tests performed. Barplots are ordered left-to-right by ascending unadjusted p-value.

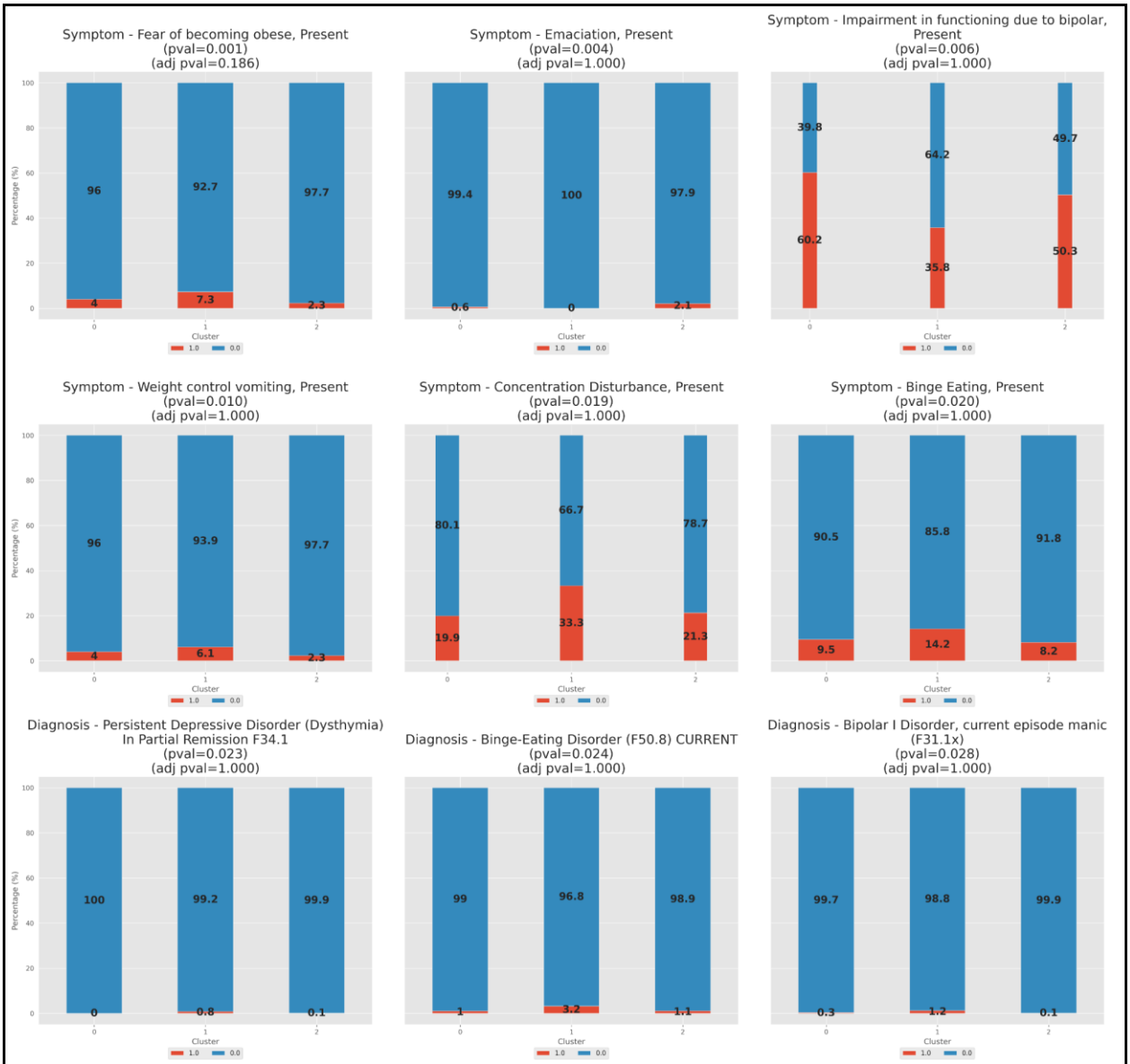


Figure 48. Year 4 outcome measure proportions per cluster[l]

[l] We report unadjusted and Bonferroni-adjusted chi-squared test p-values. TableOne package internally applies the Bonferroni correction by multiplying unadjusted p-values by the total number of tests performed. Barplots are ordered left-to-right by ascending unadjusted p-value.

4.5.2. Experiment #2: Trajectories of CBCL syndrome scale scores prior to self-injurious behavior

Clustering trajectories of the 8 CBCL syndrome scale raw scores from ages 9-13 of participants with an SIB diagnosis at year 3 (n = 310) revealed 4 distinct subgroups.

4.5.2.1. Cohort selection

We first limited our cohort to only those participants who were diagnosed with one or more of 10 KSADS self-injurious behaviors at year 3 (Table 16). We further selected based on availability of all 8 CBCL syndrome scale raw scores (Table 12) at 3 or more timepoints. After additional preprocessing and filtering, our final cohort consisted of 310 participants (Figure 49).

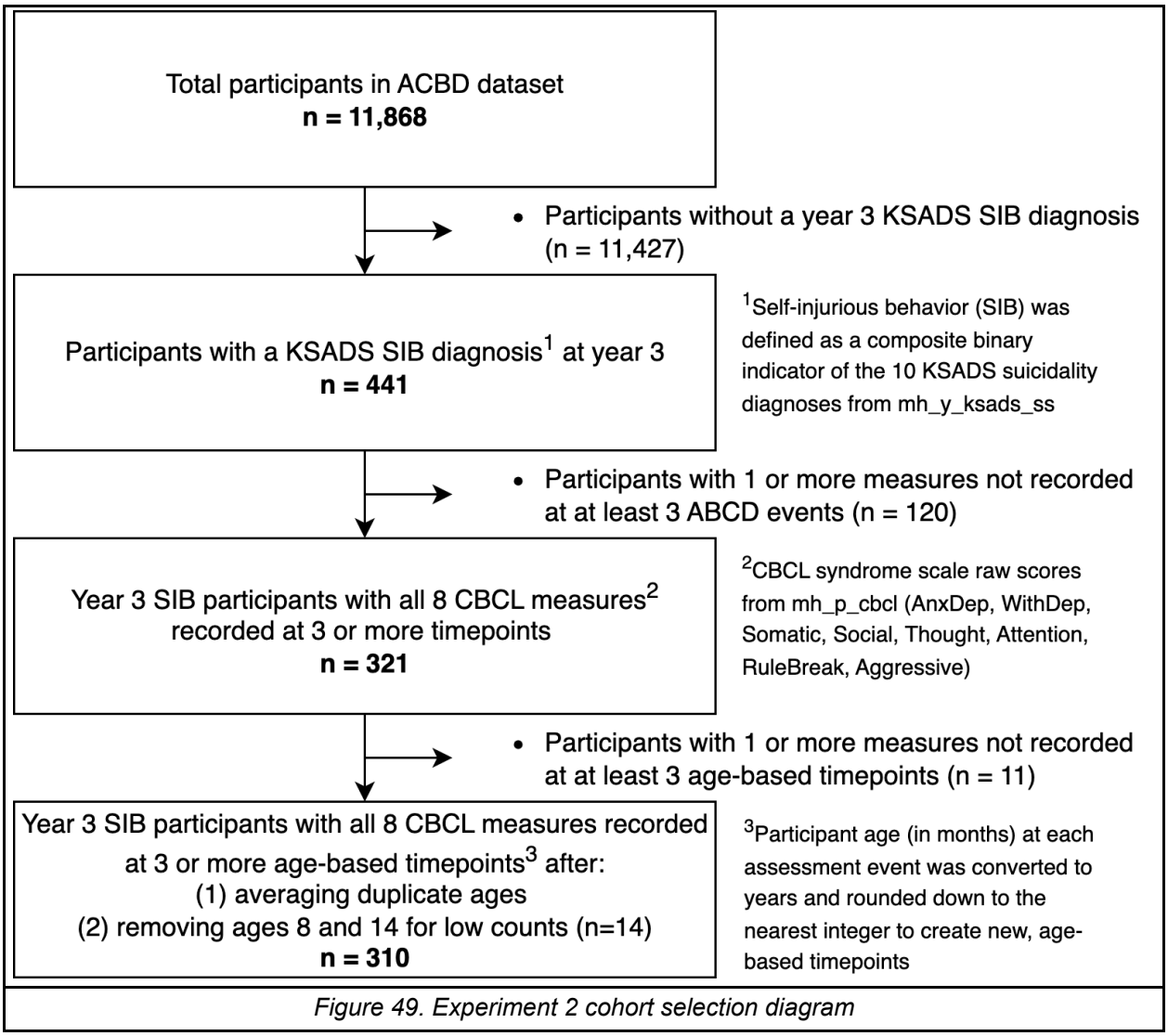


Figure 49. Experiment 2 cohort selection diagram

4.5.2.2. *Optimal cluster number selection*

We concluded that 4 clusters was optimal based on (i) reaching the elbow of the S_Dbw index curve and (ii) relative minimization (second lowest value) of the Davies-Bouldin score. Figure 50 shows calculations of each index for CRLI results specified on 2 to 6 clusters. Red points indicate the cluster number for which the index achieved the best performance.

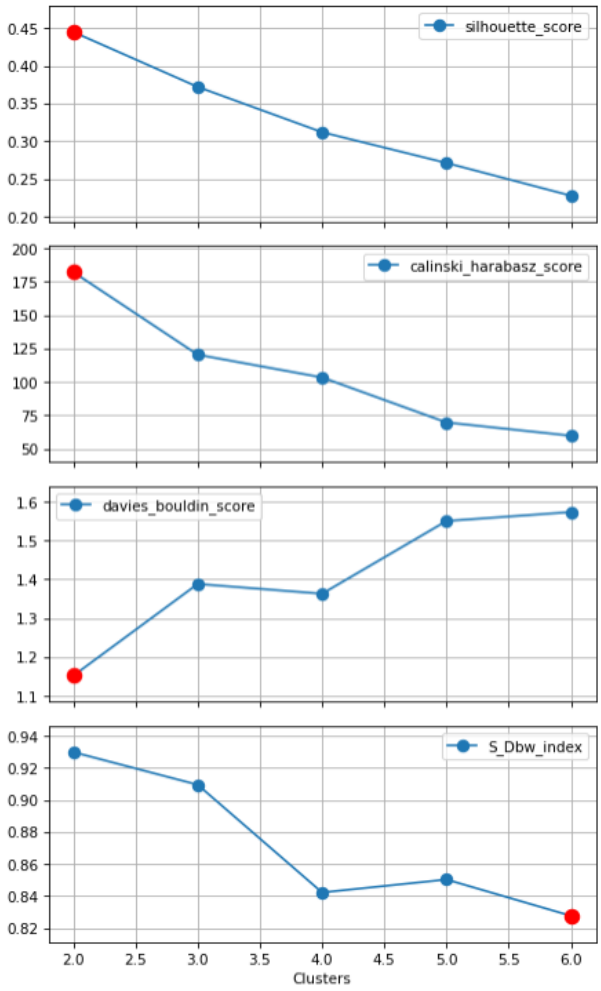
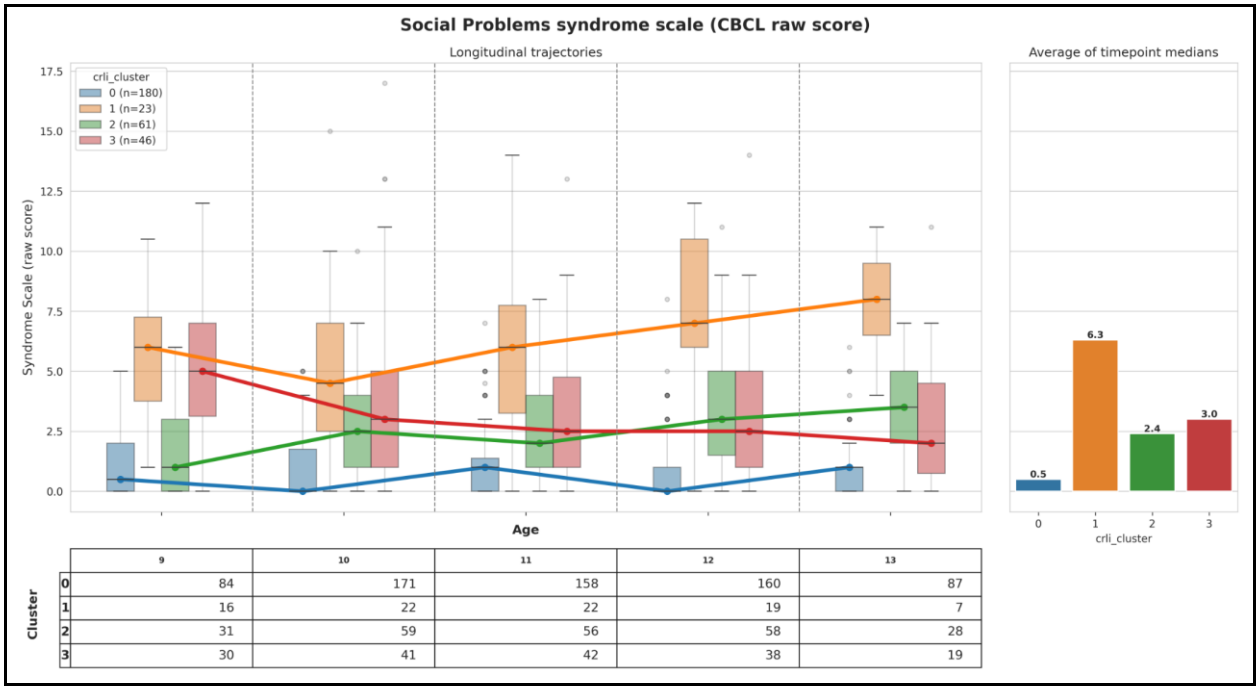
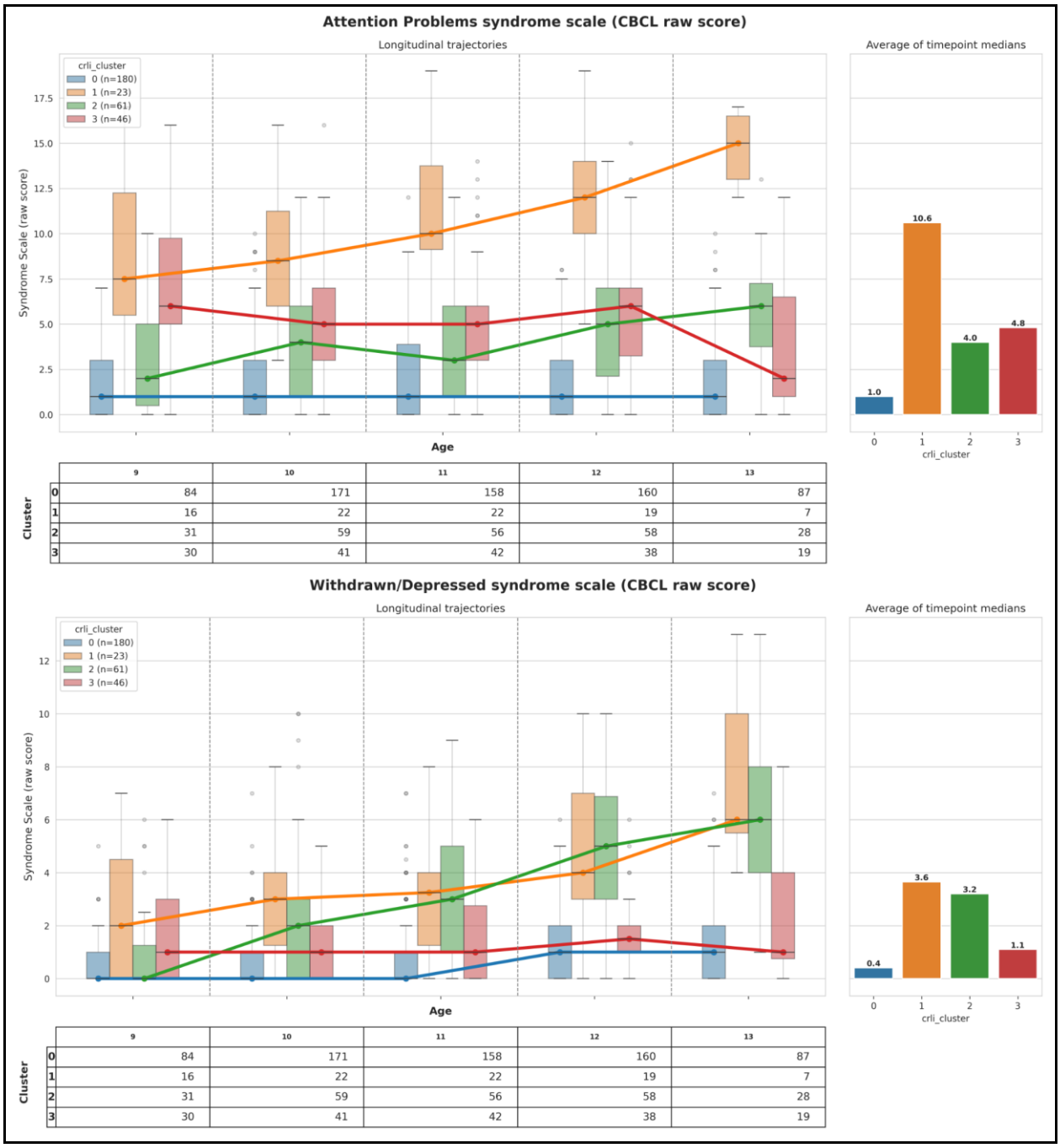


Figure 50. Internal CVI for SIB CBCL clusters

4.5.2.3. Trajectory visualization

Four of the total eight CBCL trajectories are shown in Figure 51. Cluster 1 (orange, n = 23) was characterized by a rapid increase in Attention Problems and Withdrawn/Depressed behavior, along with high levels of Social Problems and Thought Problems. Cluster 2 (green, n = 61) exhibited a similar upward trend in Withdrawn/Depressed behavior but was otherwise stable or moderately increasing. Cluster 0 (blue, n = 180) overall had the lowest and most stable scores. Lastly, Cluster 3 (red, n = 46) was similar to Cluster 2 in Social Problems and Attention Problems, similar to Cluster 0 in Withdrawn/Depressed behavior, and showed an overall longitudinal decrease in Thought Problems. A brief summary of these trends is given in Table 17. Mean lineplot overlays with 95% confidence intervals are shown in AppendixFigure 3.





Cluster	9	10	11	12	13
0	84	171	158	160	87
1	16	22	22	19	7
2	31	59	56	58	28
3	30	41	42	38	19

Cluster	9	10	11	12	13
0	84	171	158	160	87
1	16	22	22	19	7
2	31	59	56	58	28
3	30	41	42	38	19

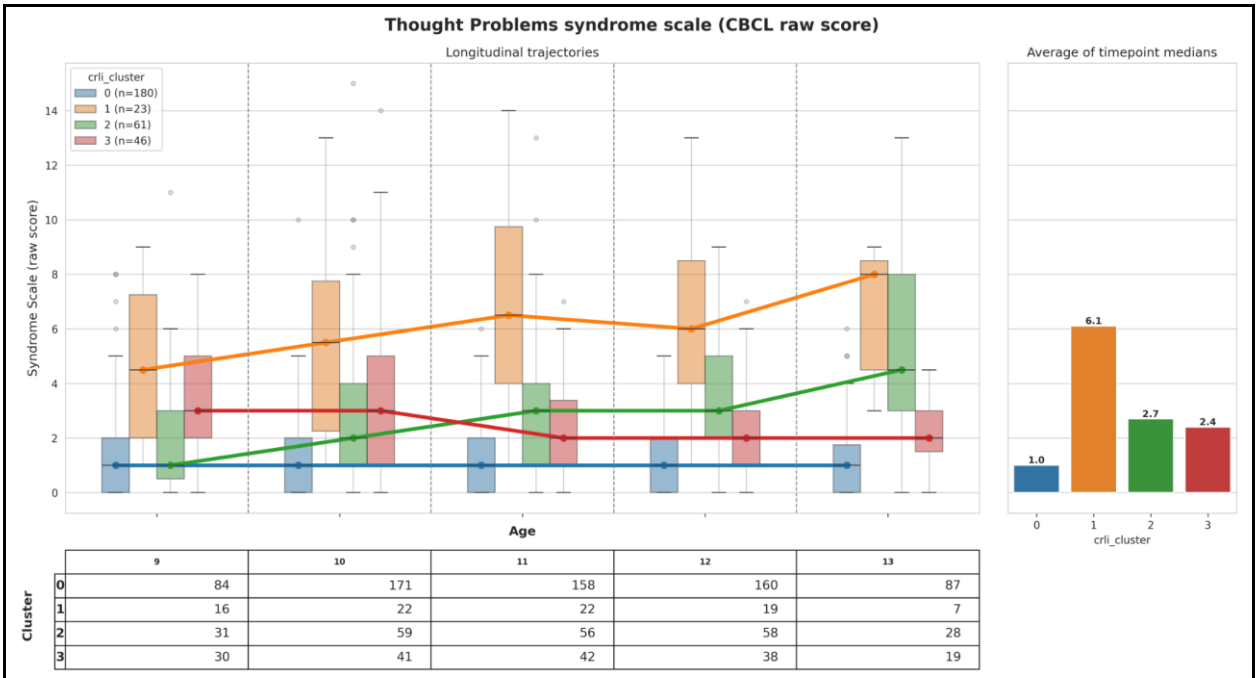


Figure 51. Multivariate CBCL trajectories of SIB-exhibiting participants

Cluster	Social	Attention	Withdrawn/Dep	Thought
0 (blue)	Low, stable	Low, stable	Low, stable	Low, stable
1 (orange)	High, increasing	High, increasing	Increasing	High, stable
2 (green)	Moderate, stable	Moderate, increasing	Increasing	Moderate, increasing
3 (red)	Decreasing	Moderate, decreasing	Low, stable	Low, stable

Table 17. Summary of Experiment #2 trajectory trends

4.5.2.4. Association of trajectories with outcome measures

Top significant associations with Year 3 KSADS outcomes reflect granular differences in distribution of types of suicidality and self-injurious behavior across clusters (Figure 52). Cluster 3 (red) had the highest proportion of self-injury that could have resulted in death (first plot in Figure) but the lowest proportion of suicidal ideation with intent to act (last plot in Figure). Cluster 2 (green) had the greatest proportion of participants ideating passively. Interestingly, Cluster 1 (orange) had lower proportions across the board, or in some cases similar proportions to Clusters 2 or 3.

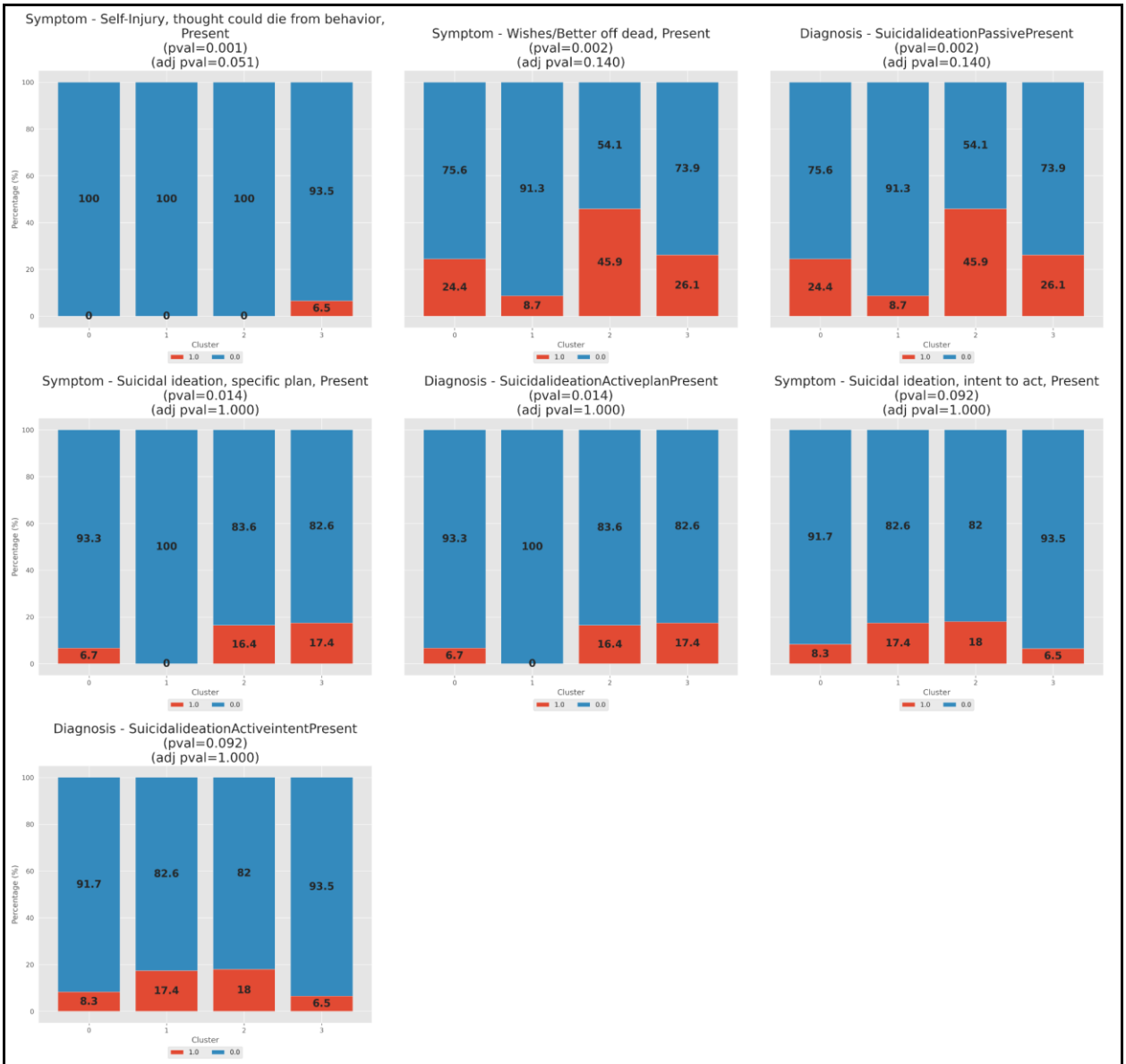


Figure 52. Year 3 outcome measure proportions per cluster[m]

Top significant associations between cluster membership and Year 4 KSADS outcomes again reflected suicidality, but also depression and anxiety (Figure 53). Cluster 1 (orange) was associated with depressive symptoms, like anhedonia (inability to experience joy) and depressed

[m] We report unadjusted and Bonferroni-adjusted chi-squared test p-values. TableOne package internally applies the Bonferroni correction by multiplying unadjusted p-values by the total number of tests performed. Barplots are ordered left-to-right by ascending unadjusted p-value.

mood, while Cluster 2 (green) was associated with anxiety and worry. Both had higher proportions of irritability. Clusters 0 (blue) and 3 (red) generally had very low proportions across all outcomes.

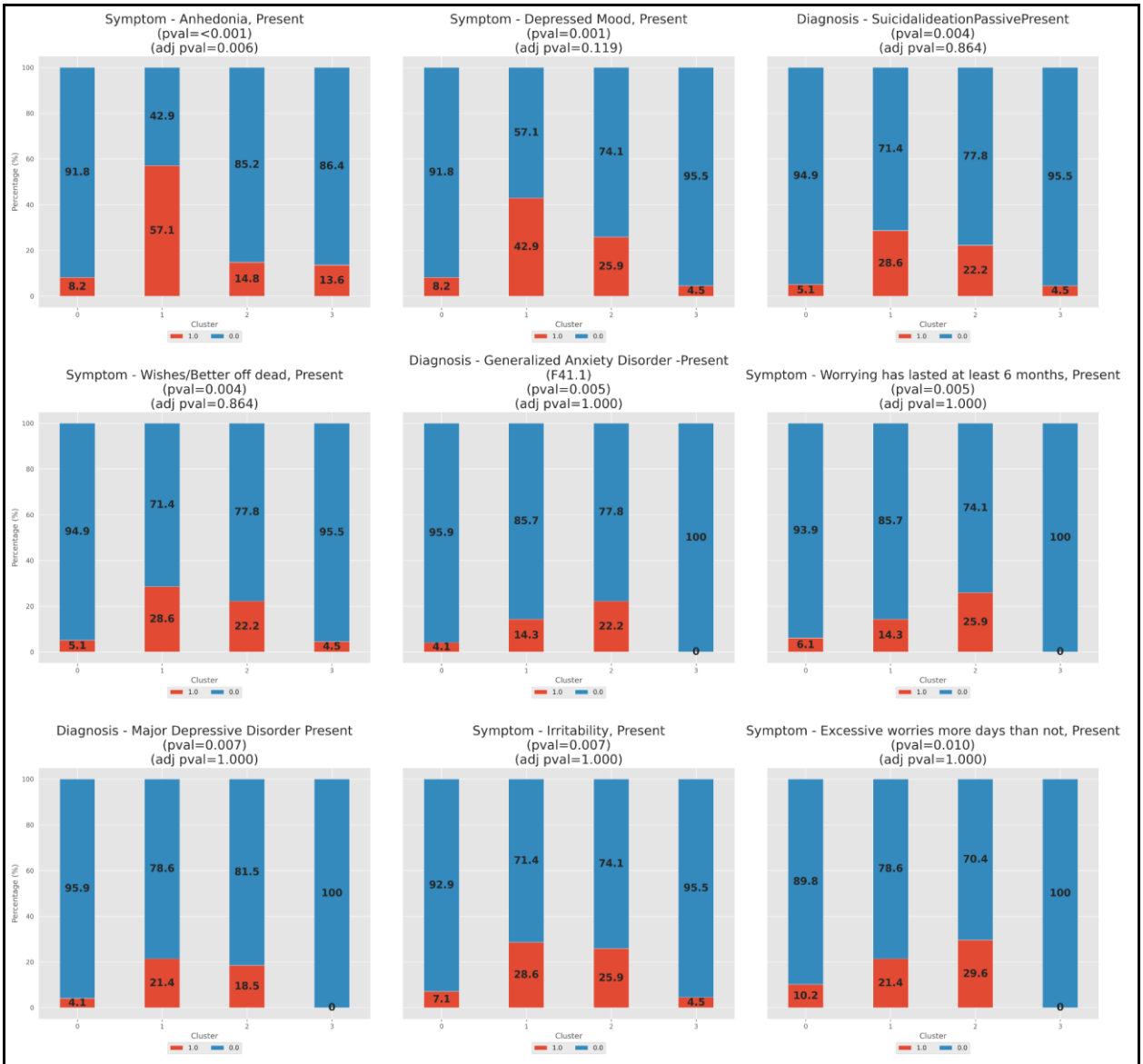


Figure 53. Year 4 outcome measure proportions per cluster[n]

[n] We report unadjusted and Bonferroni-adjusted chi-squared test p-values. TableOne package internally applies the Bonferroni correction by multiplying unadjusted p-values by the total number of tests performed. Barplots are ordered left-to-right by ascending unadjusted p-value.

4.6. Discussion

4.6.1. Conclusions

4.6.1.1. Informatics contributions

Our foremost informatics contribution is the deployment of CRLI in the context of a high-dimensional, longitudinal observational cohort. At present, CRLI is the state-of-the-art one-stage (joint optimization of imputation and clustering) incomplete time series clustering method. Thus, our demonstration of feasibility and utility in this biomedical context is an important resource for the broader research community focused on studying longitudinal trends in human health. The analytical workflow we have produced is also fully reproducible since our dataset of choice is publicly available. Additional functional contributions include comprehensive visualization strategies which help with explainability of complex trajectory patterns. In this process, we made contributions to the PyPOTS repository to uphold open-source community building.[231,232]

As discussed in Section 4.3, researchers working with ABCD and other clinical, epidemiological, and developmental datasets, have relied on more traditional, model-based statistical approaches to identify trajectory subgroups.[34] We demonstrate that CRLI, and similar deep learning-based clustering approaches, are a worthwhile consideration and may even capture trends that other methods cannot. Our work lays the foundation for further analyses as more timepoints are released in the future, including modalities that did not qualify at present due to too few repeat measurements.[233]

MVTS clustering is able to elucidate latent trajectory patterns, but only if they exist in the first place. In the process of characterizing the longitudinality of ABCD, we proposed an approach to a priori identification of longitudinal measures that may be most informative to MVTS clustering (Section 4.4.2.1). Though we did not flesh out our coefficient of variation idea completely, we offer

it as an alternative to strictly domain knowledge-based or literature-informed feature selection. By evaluating the inherent variation of each measure across all timepoints for all participants, this approach can serve as a dimensionality reduction method before applying CRLI.

4.6.1.2. Specific experimental takeaways

Through cluster identification and association with KSADS outcomes, we were able to identify protective trajectories and high-risk trajectories, which could inform clinical decision-making, screening practices, and intervention design. CRLI is able to identify specific inflection points in adolescent development where a subgroup may have an important shift in rate of change of one measure or more.

In Experiment 1, subgroups identified based on hormonal and BMI trajectories were, unsurprisingly, associated with year 4 outcomes related to weight control and eating disorders. However, we also observed differences in proportions of, on the surface, unrelated outcomes like bipolar and concentration disturbance, though these were measured in a smaller subset of the cohort (Figure 48). In Experiment 2, we saw more marked cluster differences in thematic areas like pleasure, worry, and depression, whereby some clusters had a “protective” influence against certain outcomes (Figure 53). Year 3 outcomes were overall less informative than year 4, due to a much smaller subset of youth-reported KSADS modules being administered in the former than the latter (AppendixTable 5). Nevertheless, year 3 results from Experiment 2 showed value in being able to identify specific subtypes of SIB associated with CBCL patterns (Figure 52). One possible interpretation of Cluster 3 in Experiment 2 having (a) the highest proportion of self-injury that could have resulted in death but (b) the lowest proportion of ideation with intent to act (Figure 52) is that some participants may self-injure as an alternative to ideation. However, this hypothesis requires careful analysis of the relevant KSADS lines of questioning and further causal analysis before any conclusions can be made.

4.6.2. Limitations

4.6.2.1. Cohort selection

While the ABCD sample was designed to minimize selection bias, its representativeness of the US population ultimately varies by the subset of the sample chosen for analysis and the specific measure(s) under investigation.[234] By enforcing a strict repeat measurement requirement as part of our cohort selection criteria, we may have left out participants that had informative trajectories. Though enforcing this minimum is consistent with recommendations involving developmental studies, it is nonetheless a limitation of our approach to longitudinal analysis.[63] It is also a function of the ABCD protocol, which specifies inconsistent measurement cadences between modalities.[201] As more timepoints become available through future ABCD releases, this same criteria will result in fewer excluded participants. In Aim 3, we explore the imputation ability of CRLI and VaDER to see if clusters can be identified even in the context of sparser time series.

4.6.2.2. Data preprocessing

There are several additional data preprocessing and feature engineering steps that we could have added to our workflow. We used the equation typically used for adults to calculate BMI for our cohort. For our population of children and teens, a more appropriate approach may have been to use sex-specific BMI-for-age percentiles.[185,235] Another option would have been to use height and weight as individual longitudinal variables instead of calculating BMI at all. With regard to pubertal hormone measures, we did not apply data cleaning protocols that have been outlined by others.[184]

Per the ABCD 5.1 Changes and Known issues document, Release 5.0 contained data missingness issues affecting KSADS scores, CBCL summary scores, and BPM summary scores, among others, which were updated in 5.1.[236] Though 5.1 was released after we conducted our analyses, our results may have been impacted by these data quality issues. Release 5.0 also had

a number of errors in weight and height measures. Though some may have been caught in the outlier removal step of our workflow, others may have remained and affected BMI calculations.

4.6.2.3. Clustering

Due to computational constraints, we did not perform an exhaustive search of the CRLI hyperparameter space (Table 9). Doing so may have resulted in more precise clustering results. Also, we could have selected a more comprehensive set of internal CVI on which to base our optimal cluster number selection. Besides the non-highlighted measures in Table 5, others used in related work include prediction strength and the Ray & Turi criterion.[54,68] In terms of visualization, we could have included a 2D representation of the latent space derived by CRLI to show better show cluster separation.[32,33] Lastly, our trajectory boxplots could have benefited from inclusion of reference ranges, however this was complicated by the lack of widely accepted ranges in the pediatric population, compared to adults. We posit that work like ours on a large, diverse, developmental study like ABCD can inform decision-making on how reference ranges are established.

4.6.2.4. Outcome measures

Our work only analyzed cluster associations with youth-reported KSADS measures, however inclusion of parent-reported measures could have broadened our conclusions (AppendixTable 5).[12,182] We also did not consider non-KSADS outcome measures that could also be representative of adolescent well-being, such as academic performance, sleep, and peer relationships.[237]

4.6.3. Future directions

A comparative analysis between CRLI and other multivariate trajectory subgroup identification methods, like representation learning, mixture modeling, distance-based clustering, and others,

could serve as a valuable benchmark and demonstrate pros and cons for each approach.[27,62] That said, CRLI and similar methods can be computationally expensive and not immediately digestible for the average patient, researcher, or clinician. Another useful comparison would be against simple statistical aggregate measures, like deltas, quantiles, and reference ranges. Performing more basic clustering, like k-means, on such engineered longitudinal measures could also be an interesting future study.

The significant statistical associations we identified with outcome measures warrants an investigation into the performance of predictive models using cluster membership labels as features, along with other engineered longitudinal features or single-timepoint features.[169,176] As discussed in Section 4.4.4.2, cluster labels can serve as model features and/or model outcomes. To take a step further, we can apply analytical and machine learning approaches to highly associated or predictive features to develop a more causal understanding of these relationships.[238]

In our longitudinal variable selection process, we ran into some preprocessing conundrums that could be tackled with help from a large language model (LLM). An LLM could be leveraged to (1) more confidently distinguish categorical from numerical variables, (2) identify variables across all domains that are related under a framework like The Research Domain Criteria (RDoC), and (3) given clustering results, generate nuanced trajectory summaries that are more descriptive than our offering in Table 17. A 2024 study by Ralevski et al. found that LLMs (GPT-4) could scalably and accurately identify instances of complex SDoH (housing instability) in the EHR.[239,240] While this gives us confidence that LLMs could accomplish similarly complex tasks according to our desired use cases, any LLM-generated insights would need to be validated by a domain expert.

Lastly, to fully harness the breadth and depth of ABCD, phenome-wide and genome-wide association studies could be performed to see which phenotypes, exposures, and genetic variations are associated with trajectories when many domains and instrument types are analyzed.[241–244] These results can be visualized with Manhattan plots, which convey association testing results much more efficiently than our stacked barplot approach (AppendixFigure 4). We can expand our association testing beyond outcome measures to include baseline risk factors like demographics, physical health, parent mental health, environment, traumatic life events, pregnancy exposures, family income, and more.[169]

4.7. Conclusion

In Aim 2, we evaluated CRLI's ability to identify multivariate trajectories of adolescent development in the ABCD dataset. We showed that CRLI is able to capture nonlinear trends in the context of a longitudinal observational cohort with a regular assessment cadence and high retention. Beyond the methodological takeaways, we showed that trajectory subgroups had variable associations with KSADS outcomes, emphasizing the value of person-centered approaches to understanding the individual heterogeneity inherent to adolescent maturation. Our findings indicate, like previous work in this area, that cross-sectional study of ABCD is not fully informative and that time series clustering captures trends and patterns that statistical aggregate measures do not fully reflect. Future work should compare the performance of one-stage deep clustering methods head-to-head with more traditional model-based methods, like GBTM, to validate these conclusions.

Thus far, we have explored one-stage MVTS deep clustering applicability in real-world and observational cohort contexts. Both aims reflected the immense variation that is inherent in biomedical time series data, across properties like trend, length, regularity, missingness, dimensionality, and more. We were left wondering (1) how much missingness is tolerable before latent clusters cannot be identified, (2) whether performance suffers when clusters are of greatly different sizes, and (3) how much length of time series affects discernment of distinct trajectories, among others.

In Aim 3, we sought to answer questions like these using synthetic time series data that we generate ourselves. Our goal is to systematically vary a number of time series and dataset properties in order to ascertain the limitations of VaDER and CRLI, the current state-of-the-art incomplete MVTS deep clustering methods. In the context of the dissertation, Aim 3 complements

Aims 1 and 2 by (1) formalizing those TS- and dataset-specific properties by which different MVTS clustering tasks can vary and (2) generating synthetic datasets based on variations in these properties to assess method performance under complex conditions.

5. Aim 3: Assessing the ability of CRLI and VaDER to detect trajectories in synthetic datasets under diverse data constraints

5.1. Introduction

Thus far, to better understand how and when incomplete multivariate time series (MVTS) deep clustering methods are useful in biomedical research data exploration, we have explored the ability of Clustering Representation Learning on Incomplete time-series data (CRLI) to detect meaningful trajectories in two real-world biomedical datasets, NIH *All of Us* and the Adolescent Brain Cognitive Development (ABCD) Study. In Aims 1 and 2, we observed that, despite irregular timepoint intervals, small cohort sizes, and data sparsity, CRLI was able to identify distinct clusters that had biomedical value. We also observed that MVTS clustering scenarios datasets can vary greatly across a number of (1) time series properties, like length, noise, missingness, and trend, and (2) dataset properties, like number of variables, number of clusters, and cohort size. While real-world datasets can inform our understanding of these properties, generating our own synthetic data can allow for finer-grained control over each property and a deeper understanding of each method's limitations. To date, a robust simulation-based evaluation incorporating diverse data constraints has not been conducted in the incomplete MVTS setting. In Aim 3, we assess the ability of IMVTS deep clustering methods to detect meaningful trajectories in synthetic datasets under diverse data constraints.

Though it was the focus of our earlier aims, CRLI is not the only available one-stage, incomplete MVTS deep clustering method. At the time of CRLI's publication in 2021, Variational Deep Embedding with Recurrence (VaDER), published two years earlier in 2019, was the state-of-the-art method against which CRLI was evaluated. The authors of CRLI critiqued VaDER's architectural shortcomings and demonstrated CRLI's superior performance on 7 out of 8 real-world benchmark MVTS datasets. However, in a few cases, the two methods were on par with

each other, warranting further investigation into when and why one method may outperform the other and produce clusterings more credible for downstream analysis.

In the context of the dissertation, Aim 3 complements Aims 1 and 2 by (a) formalizing those time series- and dataset-specific properties by which different MVTS clustering tasks can vary, (b) generating synthetic datasets based on variations in these properties to assess methods under complex conditions, (c) reporting VaDER and CRLI performance on comprehensive external clustering validation metrics, and (d) exploring multiple alternative 2D visualizations of learned latent representations to qualitatively understand cluster separation behavior.

5.2. Background and Significance

5.2.1. Benchmark dataset considerations

Challenging, diverse, and established benchmark datasets are crucial for the comparative evaluation of any group of machine learning methods, and clustering is no exception. To build such datasets, empirical (real-world) data from domains of interest can be sourced and used (1) as is, (2) post-modification (adding outliers, noise, missingness), or (3) as input for simulations. This last category, realistic simulations, is distinct from synthetic simulations, which are generated based on a set of carefully considered a priori technical assumptions. While some important characteristics of method performance can be derived from empirical datasets, the most important evaluation criteria cannot be assessed since we cannot isolate and manipulate the true values of the underlying parameters we are interested in. On the other hand, simulation studies allow us to investigate the behavior of methods when applied to synthetic datasets with known properties, like those summarized in Table 18. While it is impossible to predict beforehand how well a method will work on a particular dataset, simulations can at least provide systematic evidence of how methods perform on synthetic datasets with similar properties.[245] Thus, including a combination of empirical and simulated datasets when constructing a benchmark dataset archive is important for holistic evaluation of clustering methods.[246]

5.2.2. Empirical MVTS data

The UEA[o] time series archive (30 multivariate datasets, AppendixTable 6) is currently the standard consolidated benchmark for evaluating MVTS clustering methods.[247,248] There are also a handful of MVTS datasets scattered across other popular archives like (UCR, UCI[p]).[33,56,64,249,250] While these datasets are diverse across properties like time series length, cluster number, and

[o] Collaborative effort between University of East Anglia (UEA) and University of California, Riverside (UCR) researchers
[p] University of California, Irvine (UCI) Machine Learning Repository

variable number, their focus is rarely on the kinds of short, sparse (high missingness) time series that are common in many biomedical applications.[54]

5.2.3. Longitudinal biomedical dataset properties

Longitudinal datasets, and biomedical ones in particular, can vary greatly depending on a number of factors.[60,251–253] These include data source (EHR, observational cohort, clinical trial), care setting (inpatient, ambulatory), and disease prevalence (common: type 2 diabetes, 10% of US; rare: cystic fibrosis, 0.01%) and timespan (acute, chronic). As it relates to multivariate time series (MVTS) clustering, this variation manifests across a number of properties listed in Table 18.

Results from Aims 1 and 2 results reinforce this notion. Aim 1 investigated trajectories of chronic disease treatment response as measured by routinely collected measures from the EHR. Aim 2 explored trajectories of development from childhood to adolescence as measured by emotional, behavioral, and physical assessments from a longitudinal observational cohort. While the same clustering kinds of clustering methods (VaDER, CRLI) can be applied in both cases, the datasets themselves differ across the aforementioned properties in Table 18. Variation in these properties may impact clustering performance because differences in architectures between VaDER and CRLI can impact their respective abilities to cope with that variation (see Section 2.4.3 for further discussion).

To diversify the current empirical data-dominated MVTS clustering evaluation landscape, and augment it specifically with longitudinal biomedical contexts in mind, we propose a synthetic simulation approach to MVTS dataset generation that emphasizes moderate-to-high missingness

rates (50%+), short time series (≤20 timepoints), small sample sizes (<1,000), and noisy distributions.[q]

	Properties	Example medical scenario
Time series variable properties	Trend & rate of change[252]	• Increasing: BMI, cognition • Decreasing: vision, bone density • Seasonal/cyclic: hormones, body mass
	Transition(s) between stable states	• Effect of exposure (medication, environment, procedure)[60]
	Noise	• Intra- and inter-patient variability[253] • Measurement error
	Autocorrelation	• Blood sugar concentration[252]
	Length (# of timepoints)	• Acute vs. chronic[254] • Discrete vs. continuous monitoring[255]
	Missingness proportion	• Loss to follow-up[167] • Monitoring guidelines
	Missingness type (MCAR, MAR, MNAR)[57]	• Data quality • Disease severity • Ascertainment bias
Dataset properties	# of clusters	• Clinical phenotypes • Disease population subtypes[256]
	# of variables	• Clinical availability • Study protocol • Ambulatory vs. hospital-based care[257]
	# of samples (total, per cluster)	• Population/sample size[253] • Rare diseases • Typical vs. nontypical disease course • Cohort selection criteria

Table 18. MVTS data properties and corresponding biomedical scenarios

[q] See Section 2.3 for discussion of these and other longitudinal data challenges

5.3. Related work

Here we discuss the state of benchmarking for MVTS deep clustering methods, the evaluation procedures undertaken by the authors of VaDER and CRLI in their respective publications, and some existing techniques for synthetic biomedical time series generation. Our work fills a gap in the literature left by the combination of approaches that have been taken thus far.

5.3.1. Deep TS clustering benchmarking

To date, Lafabregue et al. are the only group to have conducted an extensive comparative study of deep clustering methods in the context of multivariate time series data. Notably, they broke each method down into three components, (1) architecture (type, number, and configuration of layers), (2) pretext loss, and (3) clustering loss (latent space separability), and evaluated combinations of these components (see Section 2.4.1 for discussion of these components). They applied 300 such combinations to 128 univariate datasets (UCR archive) and 30 multivariate datasets (UEA archive).[28] Though these archives were originally designed to evaluate supervised classification methods, they have been commonly used as a standard benchmark in the clustering field. While the UEA multivariate archive is robust, having sourced datasets from diverse areas of application (Human Activity Recognition, Motion classification, ECG classification, EEG/MEG classification, Audio Spectra Classification), it includes no series with missing data.[247] Thus, Lafabregue's scope did not include one-stage (imputation included) methods like VaDER and CRLI, though their analysis did include VADE and DTCR, clustering losses which VaDER and CRLI, respectively, draw from.

5.3.2. Benchmarking performed in VaDER and CRLI publications

de Jong et al. (2019) introduced VaDER and evaluated its performance on four real-world benchmark datasets (from UCI/UEA archives) and more than a hundred simulated datasets. For the synthetic portion, they used mixtures of vector autoregressive (VAR) processes to simulate

3-cluster MVTS datasets with 4 variables across 8 timepoints (AppendixFigure 5). An adjustable clusterability parameter (λ) was used to generate datasets with varying levels of separability. To assess the imputation ability of VaDER, they introduced various amounts (0-90%) and types (MCAR, MNAR)[r] of missingness and chose cluster purity as their measure of performance. Thus, while their experiments did explore missingness type and proportion, and clusterability and trend, they did not vary other properties such as number of variables, number of timepoints, number of ground truth clusters, and number of samples per cluster.[54] Furthermore, they reported only a single external cluster validation metric (purity), which is not sufficient to fully understand performance.

Ma et al. (2021) introduced CRLI and compared its performance to VaDER's on 8 real-world datasets (Table 19), but not on synthetic data in a controlled environment. They also tested several two-stage methods, first imputing by using one of three imputation methods (ZERO, GAIN, BRITS)[s], then applying one of five clustering methods (KS, DEC, IDEC, DTC, DTCR).[t] 6 of the 8 datasets (i) were univariate and (ii) had missingness ratios of 6% or less. The other 2 datasets (BloodSample, Physionet), both from the medical domain (highlighted in yellow in Table 19), had (i) the same number of true clusters (two), (ii) similarly high missingness ratios (80-86%), and (iii) larger training sizes (700+) than the other 6 datasets. Across the clustering indices the authors reported, which did not include ARI, VaDER was competitive with CRLI on these 2 sparse, multivariate biomedical datasets, especially PhysioNet. This is shown in Table 20, which is summarized from the CRLI supplement.[u] In a separate analysis, Ma et al. performed an

[r] MCAR = Missing Completely At Random; MNAR = Missing Not At Random
[s] ZERO = zero imputation; GAIN = Generative Adversarial Imputation Nets; BRITS = Bidirectional Recurrent Imputation for Time Series
[t] KS = k-Shape; DEC = Deep Embedded Clustering; IDEC = Improved Deep Embedded Clustering; DTC = Deep Temporal Clustering; DTCR = Deep Temporal Clustering Representation; MVTS imputation and deep clustering loss functions are discussed in Sections 2.4.2 and 2.4.1.3, respectively
[u] Ma et al. repeated each experiment 5 times and recorded the average results and corresponding standard deviations (shown in parentheses) as measurements for stability.

assessment of CRLI's imputation ability by randomly dropping 20% of the values of the first 20 datasets in the UCR archive (AppendixTable 7), but this analysis did not include VaDER and included different datasets than the aforementioned 8.

Dataset	Clusters	Length	Dim	Missing ratio(%)	#Train	#Test	Domain
Ali-v1	2	932	1	1.84	29	13	microarray
Ali-v2	3	1030	1	2.59	43	19	microarray
Ali-v3	4	1030	1	2.59	43	19	microarray
BloodSample	2	20	10	85.72	707	176	medical
Chen	2	2328	1	2.31	125	54	microarray
Vote	2	16	1	5.41	303	131	media
Liang	3	2505	1	0.79	25	12	microarray
Physionet	2	48	35	80.52	2798	1199	medical

Table 19. Statistics of real-world datasets used to evaluate CRLI against other methods[33]

Dataset	Metric	VaDER	CRLI	Δ (CRLI - VaDER)
BloodSample	**RI**	0.67 (0.06)	0.85 (0.02)	+0.18
	NMI	0.20 (0.00)	0.52 (0.04)	+0.32
	Purity	0.80 (0.05)	0.92 (0.01)	+0.12
	ACC	0.79 (0.05)	0.92 (0.01)	+0.13
Physionet	**RI**	0.70 (0.10)	0.76 (0.00)	+0.06
	NMI	0.03 (0.00)	0.01 (0.01)	−0.02
	Purity	0.86 (0.00)	0.86 (0.00)	0.00
	ACC	0.82 (0.11)	0.86 (0.00)	+0.04

Table 20. VaDER vs. CRLI performance on real-world multivariate datasets with high missingness[33]

5.3.3. Synthetic time series generation approaches

Several methods exist to generate generic and biomedical-specific synthetic time series. These include generative adversarial network (GAN) architectures (Wasserstein GAN, DöppelGANger), vector autoregressive processes (used in the original VaDER paper), and longitudinal plasmode simulations.[54,251,258] GAN architectures and plasmode simulations use existing data to generate synthetic, privacy-preserving data that resembles the original with regard to underlying distributional and correlation patterns. Isasa et al. developed an evaluation framework to measure

the resemblance between synthetic time series and the real time series they were generated from, but autocorrelation was the only time-series-specific metric they calculated. They mention that other time series-specific features worth checking for resemblance on include central tendency (mean, median, mode), variability metrics (range, variance), trend metrics (slope), and seasonality metrics (cycle).[251] Vector autoregressive processes, originally developed for financial analyses, are used to mathematically model variable auto- and cross-correlations between multiple time points. While these methods have been shown to generate realistic biomedical time series data, they do not allow for fine-grained control over the specific time series properties we intend to manipulate, like those mentioned by Isasa et al. and those described in Table 18.

5.3.4. Gaps in benchmarking and synthetic data Aim 3 will address

Thus far, benchmarking of incomplete MVTS deep clustering methods has been restricted to real-world and synthetic datasets with limited variability across important time series properties. Characterizing the robustness of such clustering methods to the inconsistent data landscape (Table 18) often encountered when working with biomedical datasets motivates Aim 3. Lafabregue et al. performed extensive evaluations of many methods but did not include one-stage methods, like VaDER and CRLI, in their scope. De Jong et al. generated synthetic datasets to evaluate VaDER but explored a limited set of properties. Ma et al. compared CRLI and VaDER performance on only 2 multivariate datasets with high missingness. Neither de Jong et al. nor Ma et al. reported performance on ARI in their results. Our goal is to generate many unique synthetic datasets by systematically varying a number of time series and dataset properties in order to ascertain the limitations of VaDER and CRLI, two notable one-stage methods. We believe the collection of experiments described here lays the foundation for a future, more exhaustive benchmarking framework by which any incomplete MVTS clustering method can be evaluated.

5.4. Methods

5.4.1. MVTS dataset generating process

The Python library **mockseries** grants direct control over synthetic dataset properties (numbers of clusters, variables, samples) and time series properties (trend, seasonality, noise).[259] We divide these into dataset and variable style properties (Table 21) based on how and when they are modified in our overall data generation process (Figure 55).[v] We created 20 unique MVTS datasets (Table 26) by randomly selecting 20 integers (known as scenario seeds) from {0,100} and using them as seeds to sample values uniformly from the dataset property sample spaces in Table 21. The following sections (1) describe how we used mockseries to design 7 time series variable styles with distinct shapes, (2) show visualizations of how different mockseries modifications affect time series models, and (3) report our dataset properties in full.

	Properties	**Sample spaces**
Dataset	Time series length (# of timepoints), *T*	{3, 5, 10, 15, 20, 50}
	Noise (Gaussian), ε	mean = 0 std (σ) = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
	Missingness proportion, *m*	{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}
	# of clusters, *K*	{2, 3, 5, 10, 20, 30, 50}
	# of variables, *P*	{2, 3, 4, 5, 6, 7}
	# of samples per cluster, *n*	{10, 25, 50, 100, 200, 500}
Variable style	Trend, rate of change	Discussed in following section
	Transition	
	Seasonality	

Table 21. MVTS data properties and search spaces used for dataset generation

5.4.1.1. Using mockseries to design distinct variable styles

The mockseries package is implemented as a collection of components, each of which has several types that model specific longitudinal patterns (Table 22). Each type of component is modified by a set of parameters (Table 23). For example, LinearTrend is modified by coefficient

[v] We acknowledge that some dataset properties could be modified at the variable level

(slope) and flat_base (intercept), SinusoidalSeasonality is modified by amplitude and period, and so on. Component types can be combined additively or multiplicatively to create time series models of varying shape and complexity (Figure 54).

Component	Types	Modeling situations
Trend	LinearTrend, ExponentialTrend, FlatTrend, Switch[w]	Long-run movement and discrete regime changes
Seasonality	SinusoidalSeasonality, DailySeasonality, WeeklySeasonality, YearlySeasonality	Cyclic patterns and constraint-driven daily/weekly/yearly profiles
Transition[x]	DirectTransition, LinearTransition, LambdaTransition	How a Switch moves between base and switch values (instant, linear, or custom)
Noise	GaussianNoise, RedNoise	Additive white and autocorrelated (red) noise
Interaction	AdditiveInteraction, MultiplicativeInteraction	Strategy to combine components

Table 22. Mockseries components, types, and modeling situations

Component	Type	Parameters
Trend	LinearTrend	coefficient, time_unit, flat_base
	ExponentialTrend	factor, time_unit, base
Trend	Switch	start_time, base_value, switch_value, stop_time, transition
Transition	LinearTransition	transition_window, stop_window
	LambdaTransition	transition_window, transition_function, stop_window, stop_function
Seasonality	SinusoidalSeasonality	amplitude, period, offset
Noise	RedNoise	mean, std, correlation, random_seed

Table 23. Selected component types and the parameters that modify them

[w] Switch is a component type used to move the time series from one value to another.
[x] Transition components are used to define the time window and shape of change within a Switch. See mockseries documentation for further details.

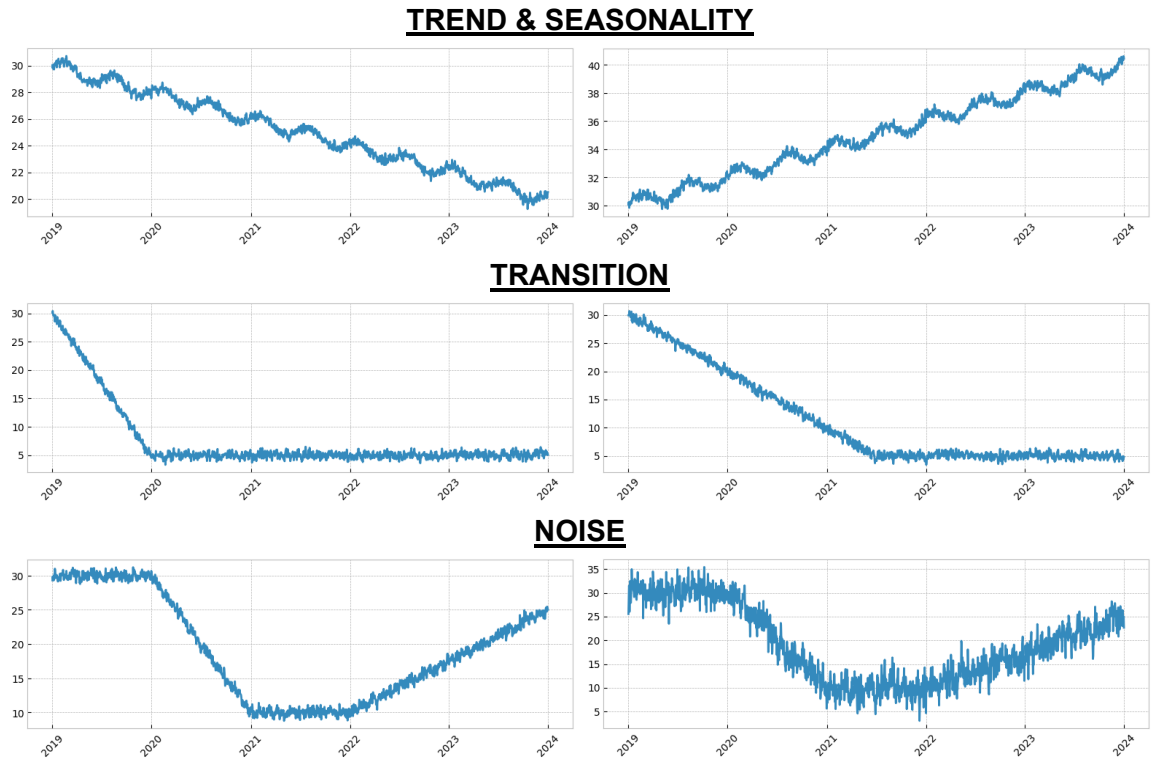


Figure 54. Mockseries component combinations yield time series models of varying shape & complexity

We define a variable style as a group of time series models that have the same mockseries component equation. We use the term variable since we are building multivariate time series datasets. We use the term time series model because we later sample from these models to populate those datasets. We constructed 5 variable styles (Table 24) inspired by longitudinal biomedical trajectories we encountered in our own work and the literature.[34,60] We combined a subset of these (Variables 0, 1, 2) to create 2 additional variable styles (Variables 5 & 6) to explicitly model cross-correlation between variables. **We utilized ChatGPT to inform parts of our variable style design (specifically, Variables 3 and 4).** In particular, we prompted ChatGPT to extend the design approach we used for Variables 0, 1, and 2 to build mockseries component equations that could generate (i) U-shaped and (ii) exponentially increasing time series models. After a few rounds of iteration, ChatGPT composed satisfactory equations for Variable styles 3 and 4, shown in Table 24.

Variable style	Description	Biomedical analogies	Mockseries component equation
Variable 0	Linear sinusoidal	Hormones; heart rate variability; gradual recovery	LinearTrend() + SinusoidalSeasonality() + RedNoise()
Variable 1	Linear transition	Acute episode; treatment response/adaptation	Switch(LinearTransition) + RedNoise()
Variable 2	Two linear transitions	Dose escalation; stepwise combination therapy	Switch(LinearTransition) + Switch(LinearTransition) + RedNoise()
Variable 3	U-shaped curve	Biomarker rebound; stress response	FlatTrend() + Switch(LambdaTransition) + Switch(LambdaTransition) + RedNoise()
Variable 4	Exponential increase	Aggressive disease; runaway metabolic process	FlatTrend() * ExponentialTrend() * [(FlatTrend() - Switch(LambdaTransition)) + Switch(LambdaTransition) * ExponentialTrend()] + RedNoise()
Variable 5	Var 0 + Var 1	--	Variable 0 + Variable 1
Variable 6	Var 0 * Var 2	--	Variable 0 * Variable 2

Table 24. Biomedical analogies for designed variable styles

As previously mentioned, each type of mockseries component is modified by a set of parameters. Since variable styles are each defined by a specific mockseries component equation, we can generate unique time series models of a given variable style by randomly sampling from search spaces of each component's parameters (Table 25).

Variable style	Mockseries component	Parameter	Sample space
Variable 0: Linear sinusoidal	LinearTrend	flat_base	{0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95}
		coefficient	{-20, -15, -10, -5, 0, 5}
		time_unit_days	{365}
	SinusoidalSeasonality	amplitude	{0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0, 6.5, 7.0, 7.5, 8.0, 8.5, 9.0, 9.5}
		period_days	{180, 270, 360, 450, 540, 630}
Variable 1: Linear transition	Switch	start_time	conditioned on *T*
		base_value	{0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95}
		switch_value	{5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19}

	LinearTransition	transition_window_days	{180, 210, 240, 270, 300, 330, 360, 390, 420, 450, 480, 510, 540, 570, 600, 630, 660, 690}
Variable 2: Two linear transitions	Switch	start_time_2	conditioned on *T*
		base_value_2	{0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95}
		switch_value_2	{5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19}
	LinearTransition	transition_window_days_2	{180, 210, 240, 270, 300, 330, 360, 390, 420, 450, 480, 510, 540, 570, 600, 630, 660, 690}
Variable 3: U-shaped curve	FlatTrend	center_level	{5, 10, 15, 20, 25, 30, 35, 40, 45}
	Switch	left_edge_level	{75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125, 130, 135, 140, 145}
		right_edge_level	{75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125, 130, 135, 140, 145}
	LambdaTransition	left_gamma	{0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9}
		right_gamma	{0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9}
		time_to_valley	conditioned on *T*
Variable 4: Exponential increase	Switch	inflection_timepoint	conditioned on *T*
	FlatTrend	start_level	{30, 35, 40, 45, 50, 55, 60}
		right_rate	{1.2, 1.21, 1.22, 1.23, 1.24, 1.25, 1.26}

Table 25. Variable style components, parameters, and parameter sample spaces

5.4.1.2. Ensuring cluster reproducibility

To create unique clusters that are different from each other across all variable styles, we ensure that the sampling behavior across all component parameters is controlled by the same seed ("cluster seed" in Figure 55). For a given MVTS dataset, we selected *K* (# of clusters) random integers from {0,100} so that each integer could act as a seed corresponding to the unique set of

component parameter values that defines that cluster's variables. Thus, one cluster seed corresponds to one distinct time series model from each variable style. Cluster seeds are consistent across datasets, meaning, for example, cluster seed 73 in datasets RandomScenario_seed-48 and RandomScenario_seed-91 generate the same component parameters, but the time series models themselves may be different due to different dataset properties selected by the scenario seed.[y] While providing a mapping of every cluster seed to its corresponding parameter values would be too space intensive, we show the mapping for RandomScenario_seed-81 in Table 28. The corresponding trajectory visualization for that dataset is shown in Figure 69. Cluster seeds are used to label the clusters they generated in figures later on.

Figure 56 shows some representative visualizations of the variety of time series models we are able to generate from each variable style with this workflow. For each variable style, each time series model shown has the same component equation but different parameter values. Those parameter values are chosen by sampling from predefined spaces (Table 25). Each color in this figure can be interpreted as a multivariate time series cluster that can be sampled from to create a dataset. The differences between the clusters can be visually understood and characterized per variable. For example, the green cluster has a flat Variable 0 and a flat Variable 1 until a linear decrease at timepoint 2030. Conversely, the red cluster has a decreasing Variable 0 and a linear transition in Variable 1 around 2024. This sampling process can be extended to any number of clusters, where each cluster has a unique seed that controls the selection of its variable style component parameters, though we show only 5 mock clusters in Figure 56.

[y] The exception to this rule is the subset of component parameters which have sample spaces conditioned on time series length, *T*, which is a dataset property set by the scenario seed. These are start_time, start_time_2, time_to_valley, and inflection_timepoint.

Of note, though we generated all 7 variable styles for each cluster in every dataset, we kept only a random subset of them, as determined by the dataset property *P* (# of variables). For each of the 20 datasets generated, the (i) scenario seed used to select dataset properties, (ii) *K* cluster seeds randomly selected from {0,100}, and (iii) *P* variable styles randomly selected from the 7 total are reported in Table 27. Our entire data generation process is diagrammed in Figure 55.

In summary, we designed a process that can generate MVTS datasets in a reproducible manner (Figure 55). Each dataset is a 3-dimensional array with axes of samples, timepoints, and variables. Variables in a dataset are unique time series models generated from 2 or more of 7 variable styles (Table 24) we designed using mockseries. Each variable style has its own mockseries component equation, which can be used to generate unique time series models by changing the parameter values of each component (Table 25). Within each dataset, clusters are distinguishable by their distinct set of parameter values across all variables (reproducibility is ensured by each cluster having its own seed). Thus, variables which come from the same style have the same overall "shape" across datasets, but can be quite different from cluster to cluster (Figure 56). Finally, an experiment/scenario is the application of both VaDER and CRLI to a single dataset, which will be discussed in coming sections.

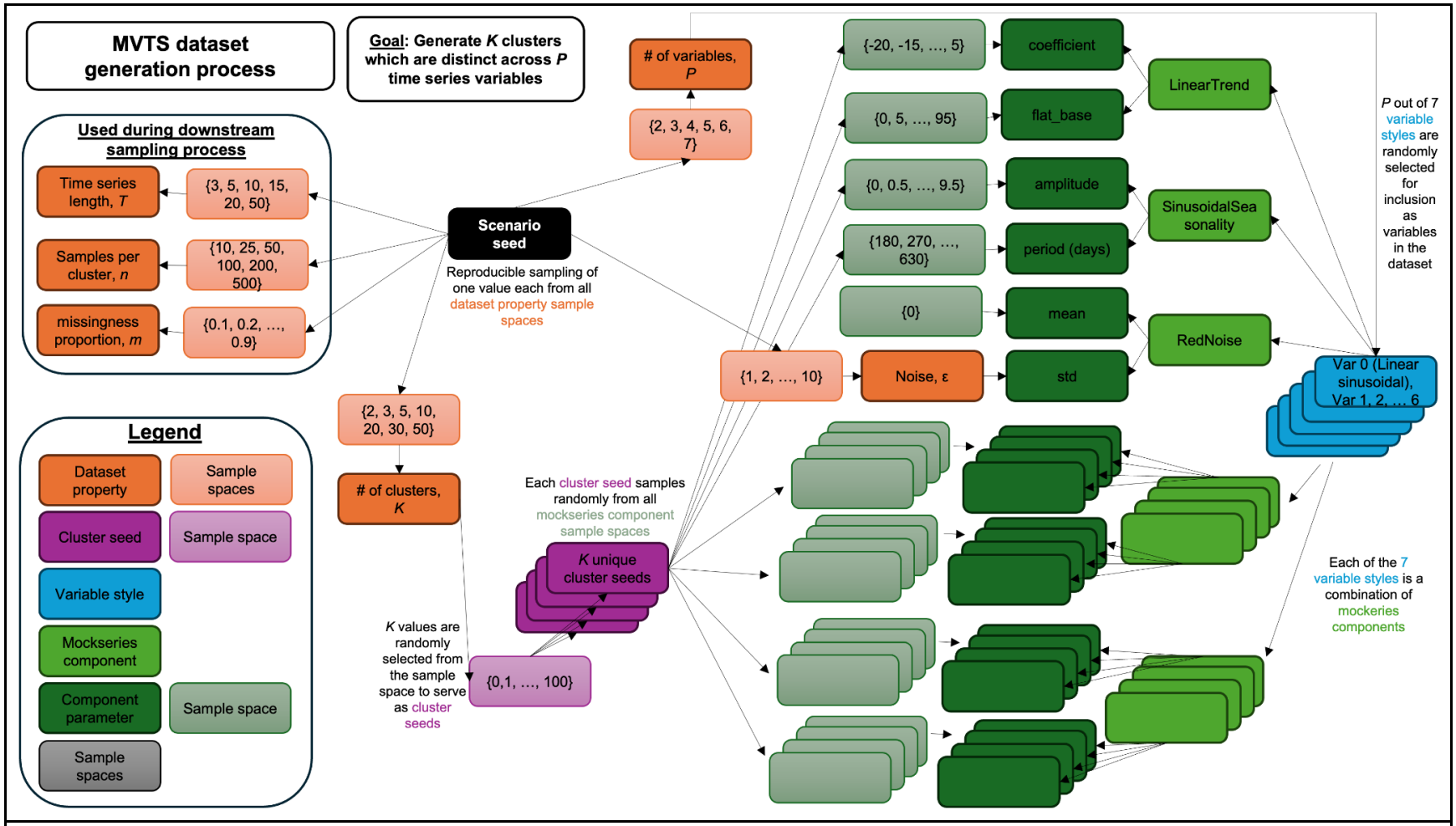


Figure 55. Dataset generation process schematic

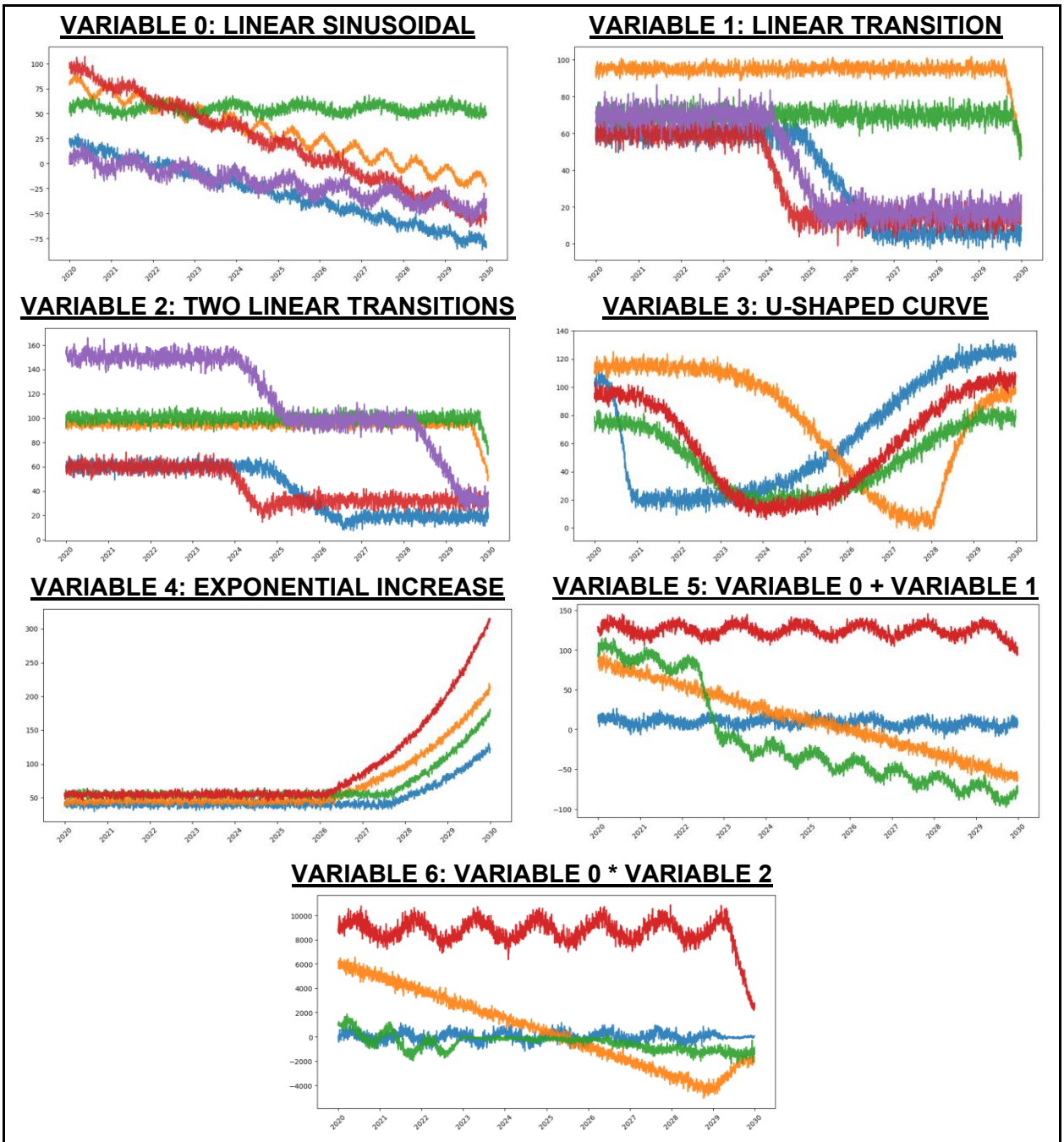


Figure 56. Unique variable styles generated using parameter sampling approach

5.4.1.3. Sampling from models to create scenario datasets

After generating a set of time series models (per variable style, per cluster), we sampled from each model n times at T timepoints to create MVTS datasets with ground truth cluster labels. We used cluster seeds directly as cluster labels to maintain consistency across scenarios. Figure 57 shows an example of this sampling behavior with T = 10, n = 100, K = 5, and cluster seeds = {38, 73, 76, 71, 69} for Variables 0 and 1 (left and right columns, respectively).

5.4.1.4. Missingness injection

For each sample array, the number of indices selected for replacement with missing is determined by the missingness proportion, m, and the length of the time series, T. We randomly select $m*T$ (truncated to integer) indices from each array and set them to NA. Figure 58 shows an example of this behavior using the same samples from Figure 57 (first row) with m = 0.7 (second row). In the second row, the original sample time series are shown faintly in gray, while the post-missingness time series are shown in their respective cluster colors, demonstrating the proportion of values (70%) that were dropped.

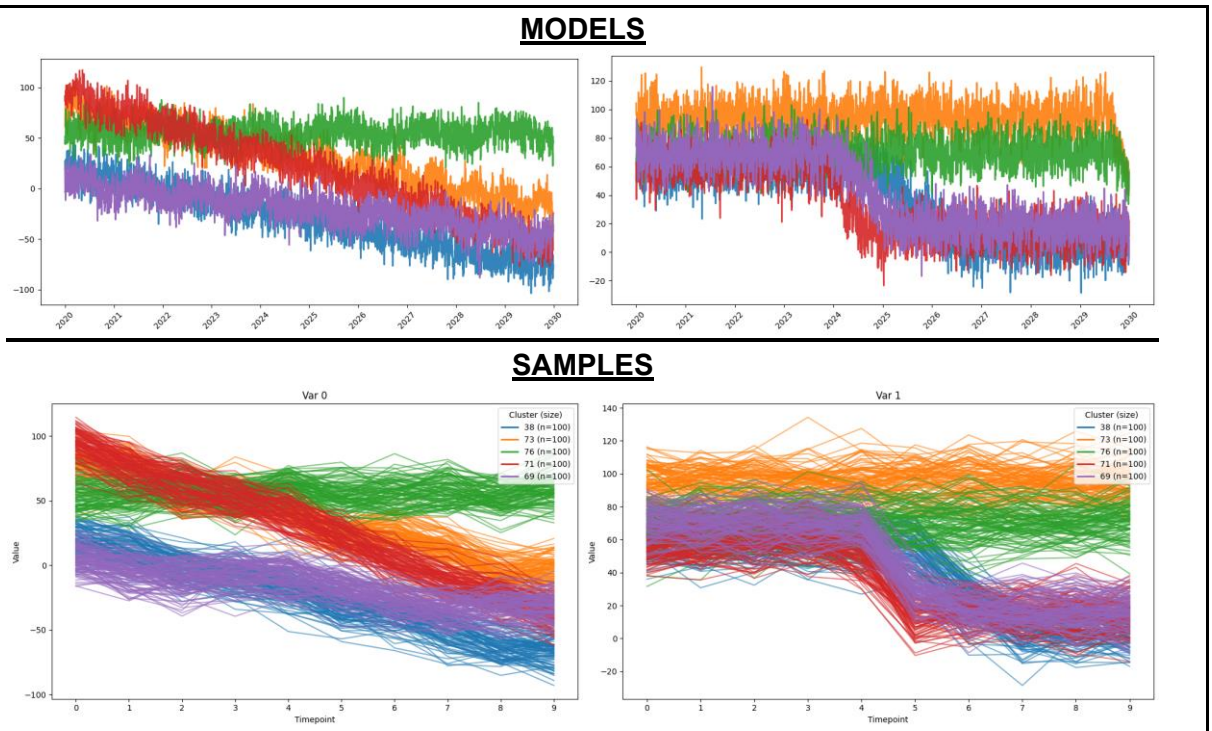


Figure 57. Mockseries sampling behavior

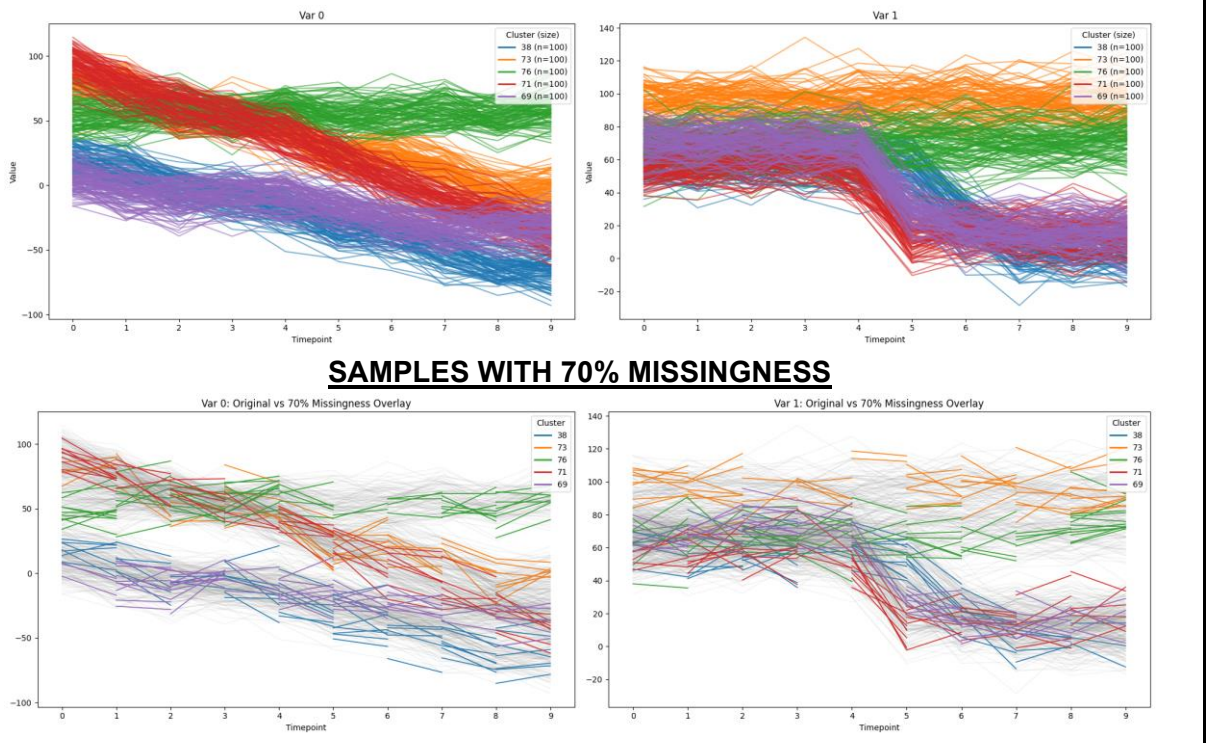


Figure 58. Visualization of 70% random missingness injection

5.4.1.5. Dataset properties

The following set of figures show how the modification of dataset properties affects time series models for a single variable style (Variable 0, though the modifications apply to all variable styles similarly).

Figure 59 shows the modification of *K*, number of clusters. Each time series model shown corresponds to a cluster seed which is what is used to sample from the mockseries component parameter spaces to generate unique time series models within the same variable style. As seen, though some time series models are similar, none of them are the same, since they have all been generated with different cluster seeds.

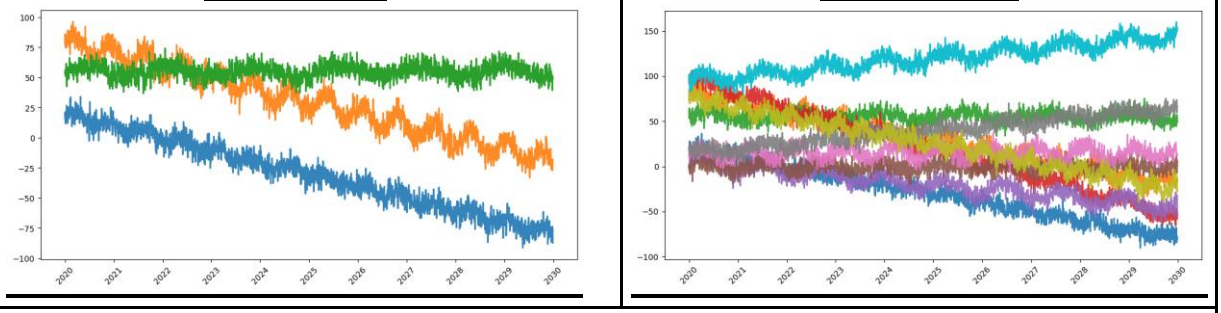


Figure 59. Changing number of clusters, K (3 vs. 10)

Figure 60 shows the modification of *T*, time series length. Despite the same cluster seeds, meaning the same underlying time series models, in both images, the clusters "look" very different because there is more temporal information available in the image on the right. Thus, despite having the same cluster seeds, datasets can end up being very different according to changes in dataset properties. Clusters 39 (green) and 50 (orange) have low separability from timepoints 0-8, but from 9 onwards they begin to diverge.

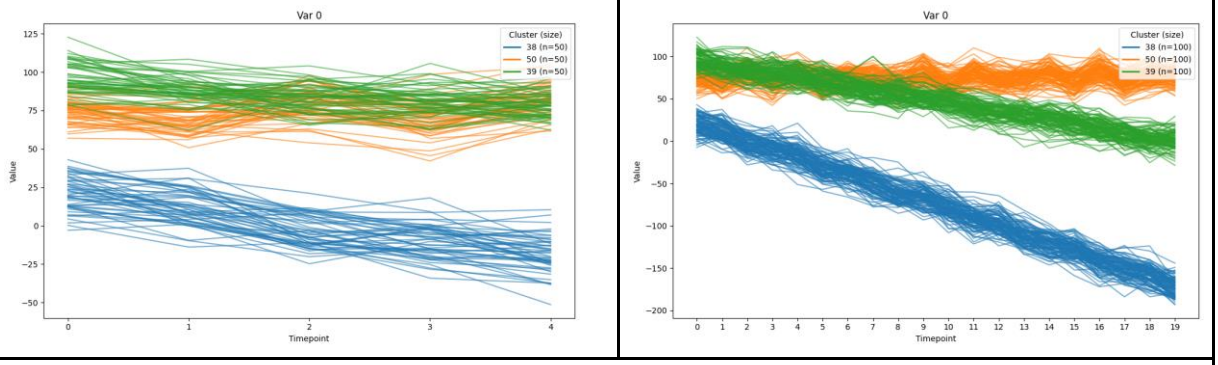


Figure 60. Changing number of timepoints, T (5 vs. 20)

Figure 61 shows the modification of σ, which controls the noisiness of the time series models. Increasing noise past a certain point can affect the separability of the final clusters. In the below figure, the red, purple, and blue time series models are easier to distinguish from 2026-2028 in the lower noise environment on the left than the higher noise environment on the right. Combined with a lower sample size, higher noise can make it difficult to identify underlying clusters.

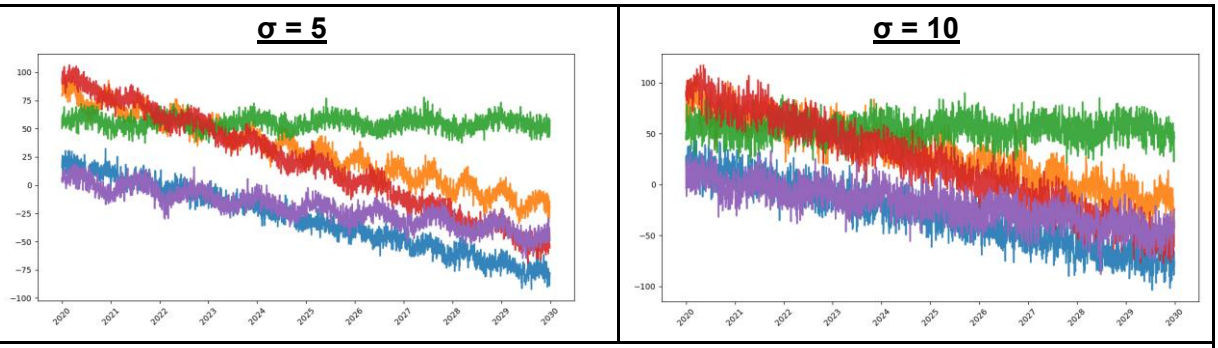


Figure 61. Changing amount of noise, σ (5 vs. 10)

Figure 62 shows the modification of *n,* number of samples per cluster. In practice, clusters may not be the same size (imbalanced), however in our experiments, *n* was the same for each cluster within a given dataset (balanced). This figure shows how increasing sample size can work to counteract noisy time series models. Despite the same noisiness in both images, the clusters with more samples (on the right) are visually easier to distinguish. Given more samples (training set), methods may perform better, since they have more data points to learn patterns from.

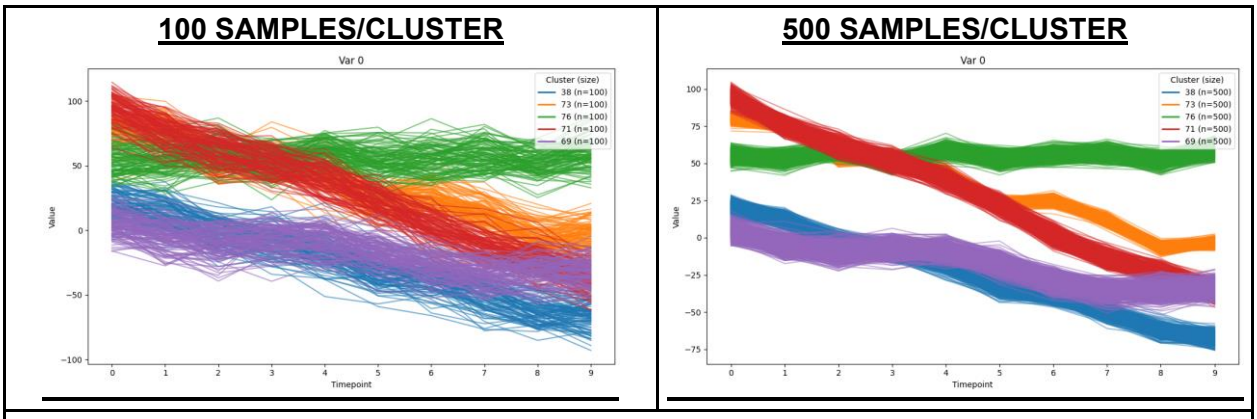


Figure 62. Changing number of samples per cluster, n (100 vs. 500)

5.4.1.6. Summary of generated dataset properties

We generated a total of 20 datasets which spanned a variety of time series and dataset properties. Reported below, per scenario, are number of timepoints, number of variables, number of clusters, number of samples per cluster, noise (σ parameter), and missingness proportion (Table 26). We also included train/test splits, which will be explained in the following section. The specific variable styles each scenario was built around and the seeds used to generate unique clusters from those styles are reported in Table 27. As mentioned earlier, since providing a mapping of every cluster seed to its corresponding parameter values would be quite space intensive, we show the mapping for RandomScenario_seed-81 in Table 28 as a representative example.

Dataset	timepoints, *T*	vars, *P*	clusters, *K*	samples per cluster, *n*	train set size	test set size	noise, ε	missingness, *m*
RandomScenario_seed-15	50	3	30	50	1050	450	7	0.7
RandomScenario_seed-23	3	3	20	10	140	60	7	0.4
RandomScenario_seed-26	50	4	3	10	21	9	5	0.5
RandomScenario_seed-30	3	4	10	10	70	30	3	0.7
RandomScenario_seed-31	15	5	2	200	280	120	10	0.4
RandomScenario_seed-32	50	4	20	50	700	300	2	0.8
RandomScenario_seed-4	20	7	10	500	3500	1500	10	0.8
RandomScenario_seed-44	20	5	3	50	105	45	2	0.8
RandomScenario_seed-48	3	6	20	100	1400	600	4	0.4
RandomScenario_seed-53	20	6	30	25	525	225	1	0.6
RandomScenario_seed-59	20	5	30	10	210	90	6	0.7
RandomScenario_seed-6	10	7	5	50	175	75	6	0.5
RandomScenario_seed-62	10	6	30	200	4200	1800	7	0.5
RandomScenario_seed-67	20	4	2	100	140	60	5	0.5
RandomScenario_seed-79	3	7	2	500	700	300	10	0.8
RandomScenario_seed-81	3	5	3	500	1050	450	5	0
RandomScenario_seed-88	15	6	20	200	2800	1200	3	0.9
RandomScenario_seed-91	20	2	50	200	7000	3000	3	0.3
RandomScenario_seed-94	3	3	30	200	4200	1800	3	0.9
RandomScenario_seed-96	20	3	30	100	2100	900	8	0.8

Table 26. Synthetic MVTS dataset properties

Dataset	variable styles	cluster seeds
RandomScenario_seed-15	[2, 3, 5]	[55, 90, 6, 73, 2, 75, 33, 39, 53, 24, 23, 14, 74, 67, 72, 7, 41, 78, 42, 94, 99, 50, 93, 65, 18, 35, 34, 80, 100, 98]
RandomScenario_seed-23	[0, 2, 6]	[10, 29, 34, 6, 40, 76, 22, 70, 87, 93, 55, 90, 73, 2, 75, 33, 39, 53, 24, 23]
RandomScenario_seed-26	[1, 3, 5, 6]	[87, 93, 55]
RandomScenario_seed-30	[2, 3, 4, 5]	[69, 48, 44, 22, 42, 45, 80, 24, 56, 4]
RandomScenario_seed-31	[0, 1, 2, 3, 5]	[88, 12]
RandomScenario_seed-32	[0, 1, 3, 6]	[80, 57, 84, 29, 56, 68, 55, 17, 45, 43, 33, 88, 1, 89, 48, 77, 81, 13, 22, 79]
RandomScenario_seed-4	[0, 1, 2, 3, 4, 5, 6]	[45, 12, 55, 99, 51, 7, 25, 65, 30, 0]
RandomScenario_seed-44	[2, 3, 4, 5, 6]	[29, 14, 34]
RandomScenario_seed-48	[0, 1, 2, 3, 4, 6]	[6, 43, 68, 4, 53, 32, 64, 36, 66, 52, 50, 84, 23, 39, 72, 47, 74, 3, 38, 73]
RandomScenario_seed-53	[0, 1, 2, 3, 5, 6]	[2, 23, 64, 44, 81, 8, 43, 12, 21, 80, 88, 76, 55, 34, 22, 5, 9, 24, 45, 15, 27, 68, 32, 62, 95, 99, 100, 35, 93, 60]
RandomScenario_seed-59	[0, 1, 4, 5, 6]	[97, 9, 49, 85, 83, 67, 24, 90, 20, 10, 6, 48, 43, 30, 93, 66, 95, 19, 25, 71, 46, 76, 0, 84, 41, 2, 14, 64, 86, 8]
RandomScenario_seed-6	[0, 1, 2, 3, 4, 5, 6]	[55, 90, 6, 73, 2]
RandomScenario_seed-62	[0, 1, 2, 3, 4, 5]	[87, 95, 60, 98, 93, 8, 82, 91, 45, 14, 77, 33, 15, 2, 23, 64, 44, 81, 99, 43, 12, 21, 76, 55, 34, 22, 5, 9, 24, 92]
RandomScenario_seed-67	[0, 1, 2, 6]	[87, 95]
RandomScenario_seed-79	[0, 1, 2, 3, 4, 5, 6]	[44, 15]
RandomScenario_seed-81	[0, 1, 3, 4, 6]	[97, 9, 49]
RandomScenario_seed-88	[0, 1, 2, 4, 5, 6]	[83, 65, 18, 35, 34, 85, 88, 23, 55, 6, 43, 68, 4, 53, 32, 64, 36, 66, 52, 50]
RandomScenario_seed-91	[1, 5]	[44, 15, 94, 67, 23, 4, 30, 48, 79, 95, 26, 90, 62, 81, 93, 6, 53, 59, 32, 31, 29, 14, 34, 8, 72, 20, 22, 40, 7, 3, 63, 75, 35, 5, 88, 71, 68, 0, 13, 73, 51, 28, 54, 42, 92, 83, 78, 27, 77, 74]
RandomScenario_seed-94	[3, 4, 5]	[87, 95, 60, 98, 93, 8, 82, 91, 45, 14, 77, 33, 15, 2, 23, 64, 44, 81, 99, 43, 12, 21, 76, 55, 34, 22, 5, 9, 24, 92]
RandomScenario_seed-96	[1, 2, 5]	[76, 22, 70, 87, 93, 55, 90, 6, 73, 2, 75, 33, 39, 53, 24, 23, 14, 74, 67, 72, 7, 41, 78, 42, 89, 85, 50, 88, 65, 18]

Table 27. Synthetic MVTS dataset variable styles and cluster seeds

	Cluster seeds ->	**97**	**9**	**49**
Variable 0	flat_base	10	40	0
	coefficient	0	5	-10
	time_unit_days	365	365	365
	amplitude	0.5	9.5	8.5
	period_days	360	270	450
Variable 1	start_time	2021-07-24 0:00:00	2021-12-21 0:00:00	2021-07-24 0:00:00
	base_value	95	75	60
	switch_value	15	14	18
	transition_window_days	270	540	510
Variable 2	start_time_2	2022-05-20 0:00:00	2022-11-16 0:00:00	2022-05-20 0:00:00
	base_value_2	70	90	0
	switch_value_2	7	18	12
	transition_window_days_2	180	630	570
Variable 3	time_to_valley	2	2	1
	center_level	5	45	45
	left_edge_level	90	75	75
	right_edge_level	100	75	110
	left_gamma	1.9	1.6	0.7
	right_gamma	1.5	1.1	0.6
Variable 4	inflection_timepoint	547.5	547.5	547.5
	start_level	40	55	35
	right_rate	1.26	1.23	1.26
Noise[z]	mean	0	0	0
	std	5	5	5

Table 28. RandomScenario_seed-81 variable style component parameters (K = 3)

[z] As a reminder, noise is set at the dataset level by the scenario seed, so it is constant across all variables and clusters for a given dataset.

5.4.2. Testing of clustering methods on synthetic datasets

We applied VaDER and CRLI to each of our datasets to see how well they could recapitulate ground truth clusters.[33,54] At the time of writing, these methods are the state-of-the-art one-stage (joint imputation and clustering) MVTS deep clustering approaches. Due to computational constraints, we ran both clustering methods with default hyperparameters (Table 29). We set the number of clusters in the clustering task to the ground truth *K* (Table 26) for each scenario. We used the PyPOTS implementations of both methods, and hyperparameter descriptions can be found in the PyPOTS github repository.[143,146,260] Full datasets were split 70/30 intro training and test sets at the times series level and stratified on cluster membership, meaning that each entire time series was either in the training or test set. For example, under this paradigm, a dataset with 2 clusters (A, B) and 100 samples per cluster would be split into a train set of 140 (50% A, 50% B) and a test set of 60 (50% A, 50% B). This ensured that both sets were proportionally identical with regard to cluster representation (Table 26).

PyPOTS hyperparameter	Description	VaDER	CRLI
Pretraining epochs	number of epochs for pretraining the model (reconstruction loss only, ignoring clustering loss)	100	--
Training epochs	number of epochs for training the model (joint loss function)	100	100
Learning rate	controls step size taken during each iteration when updating model parameters	1e-4	5e-3
Hidden layers[aa]	# of layers of the encoder: [# of units per layer]	1: [64]	1: [50]
Batch size	number of samples processed together in a single training iteration	16	16
d_mu_stddev	dimension of the mean and standard deviation of the Gaussian distribution (VaDE latent space)	10	--
lambda_kmeans	Coefficient of the soft K-means objective (encourages learned representations to form cluster structures)	--	1e-3
n_clusters	number of clusters in the clustering task	Ground truth, *K*	

Table 29. VaDER and CRLI hyperparameters

[aa] The VaDER repository default is 2: [12,2] and the default mentioned in the paper is 2: [36,4]. We chose 1: [64] to match the default CRLI architecture while still being in the hyperparameter space that was explored in the VaDER paper.

5.4.2.1. Variational Deep Embedding with Recurrence (VaDER)

VaDER was published by de Jong et al. in 2019. The architecture, shown in Figure 29, combines variational autoencoder principles with long short-term memory (LSTM) networks and implicit imputation. This method was developed with multivariate clinical patient trajectories in mind and was originally evaluated on simulated data and benchmark classification datasets.[54]

5.4.2.2. Clustering Representation Learning on Incomplete time-series data (CRLI)

As discussed in Section 3.4.4., CRLI, published by Ma et al. in 2021, is another deep learning method which simultaneously performs missing value imputation and clustering across MVTS. Notably, Ma et al. compared their method's performance against VaDER's and critiqued VaDER for (a) not sufficiently preventing error propagation from imputation to clustering and (b) requiring a relatively large amount of training data (training size > 700, Table 19) to perform on par with CRLI (see Section 2.4.3.3 for further discussion).[33]

5.4.3. Performance evaluation

To evaluate the model-predicted clusters against ground truth, we used four external cluster validation indices (CVI): cluster purity, Rand Index (RI), Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI). External CVI are typically classified into three categories according to their similarity assessment technique: pair-counting (RI, ARI), information theoretic (NMI), and set-matching (purity).[261] These are used widely in the clustering literature to quantify clustering method performance and have been described in detail elsewhere (see Section 2.4.4.2).[27] Purity measures how pure each predicted cluster is with respect to the true class labels. RI is the proportion of point pairs that are either in the same cluster in both partitions or in different clusters in both. ARI is a corrected-for-chance extension of the RI. And, lastly, NMI evaluates the correlation (mutual information) between two clusterings.[27] Purity, RI, and NMI values range from 0 to 1, while ARI ranges from -1 to 1, with 1 indicating an identical match between predicted and

ground truth clusters. The equations for each of these metrics is included in Table 6 and all were implemented in PyPOTS. We calculated all 4 metrics for both methods (VaDER, CRLI) on the held out test set in each of the 20 datasets. We also measured runtime, in minutes, for each scenario using Python's time.time() method.

For discussion on the differences between external and internal CVI, the latter of which are used to evaluate clustering performance in the absence of ground truth labels, see Section 3.4.5

5.4.4. Visualization strategies

Following others, we applied principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP) to map the test set latent representations learned by VaDER and CRLI into two dimensions for easier visualization.[28,32,33] These methods are used widely for dimensionality reduction and visualization of high dimensional datasets. While PCA is a linear method that struggles to perform in cases where relationships are non-linear, t-SNE and UMAP are non-linear manifold learning methods which find local and global structures with varying efficacy. They have been described in detail elsewhere.[262,263]

5.5. Results

We first report test set performance metrics and runtimes for both methods across all scenarios in Table 30. This can be referenced against Table 26 in the Methods section to compare how performance was impacted by differences in dataset properties. We then analyze five scenarios in detail to (a) visualize our methodological workflow, (b) show examples of varying clustering difficulty, and (c) highlight differential performance between VaDER and CRLI. Though we report multiple clustering validation indices and 2D visualizations for each experiment, we base our analysis of method performance primarily on (1) Adjusted Rand Index results, as discussed in Section 2.4.4.2, and (2) UMAP visualizations, which have been shown to outperform t-SNE in the deep clustering context.[263]

5.5.1. Full performance results

In Table 30 we report Purity, Rand Index (RI), Adjusted RI (ARI), Normalized Mutual Information (NMI), and runtime (in minutes) per method for each dataset we constructed. We show many instances where the non-ARI measures exaggerated method performance (examples: RandomScenario_seed-23, RandomScenario_seed-53, RandomScenario_seed-94). This is especially true for RI, despite its popularity in prior studies. ARI, and adjusted measures in general, are more suitable for clustering since they are independent of the number of clusters.[89] Overall, ARI crossed 0.5 in only 4/40 cases: CRLI on RandomScenario_seed-26, 31, and 44 and VaDER on RandomScenario_seed-91. We propose that (1) use of non-ARI metrics has inflated the perception of method performance in general and (2) performance is generally poor across most datasets with a few notable exceptions. We found that runtime increases were generally a function of (1) time series length, T and (2) train set size.

In Table 31 we report performance deltas (CRLI − VaDER). Positive values indicate CRLI > VaDER (green); negative values indicate VaDER > CRLI (blue). We show that CRLI outperforms VaDER on 17/20 datasets for ARI. However, the magnitude of the difference in ARI between the two methods is less than 0.1 for 6/20 datasets and less than 0.2 for 11/20 datasets. Thus, CRLI's performance on ARI is only marginally better than VaDER across the majority of datasets. This is in line with the evaluation conducted by Ma et al. in the CRLI publication (Table 20). This demonstrates that (1) performance is highly dataset dependent across both methods but (2) there is a subset of datasets for which one method outperforms the other.

Dataset	Method	Purity	RI	ARI	NMI	Runtime (mins)
RandomScenario_seed-15	CRLI	**0.427**	**0.945**	**0.247**	**0.619**	206.5
	VaDER	0.309	0.893	0.111	0.499	**53.3**
RandomScenario_seed-23	CRLI	**0.383**	**0.829**	0.007	**0.540**	0.2
	VaDER	0.183	0.728	**0.014**	0.318	0.8
RandomScenario_seed-26	CRLI	**0.889**	**0.861**	**0.643**	**0.786**	2.7
	VaDER	0.667	0.639	0.071	0.393	1.5
RandomScenario_seed-30	CRLI	0.433	0.851	0.031	0.566	0.6
	VaDER	**0.500**	0.851	**0.078**	**0.575**	0.4
RandomScenario_seed-31	CRLI	**1.000**	**1.000**	**1.000**	**1.000**	7.5
	VaDER	0.558	0.503	0.007	0.013	5.7
RandomScenario_seed-32	CRLI	**0.410**	**0.924**	**0.216**	**0.555**	61.7
	VaDER	0.257	0.902	0.059	0.344	35.6
RandomScenario_seed-4	CRLI	**0.413**	**0.854**	**0.225**	**0.424**	127.7
	VaDER	0.243	0.827	0.120	0.300	75.5
RandomScenario_seed-44	CRLI	**0.867**	**0.855**	**0.675**	**0.765**	7.6
	VaDER	0.644	0.652	0.240	0.309	4.5
RandomScenario_seed-48	CRLI	**0.243**	0.875	**0.059**	**0.294**	10.2
	VaDER	0.155	**0.877**	0.017	0.151	7.8
RandomScenario_seed-53	CRLI	**0.360**	**0.938**	**0.123**	**0.558**	19.2

	VaDER[bb]	0.222	0.922	0.035	0.441	13.8
RandomScenario_seed-59	CRLI	**0.411**	**0.948**	**0.037**	**0.677**	7.9
	VaDER	0.278	0.901	0.004	0.538	4.7
RandomScenario_seed-6	CRLI	**0.613**	**0.799**	**0.393**	**0.546**	3.3
	VaDER	0.347	0.697	0.027	0.123	2.1
RandomScenario_seed-62	CRLI	**0.443**	**0.951**	**0.284**	**0.580**	80.3
	VaDER	0.103	0.908	0.013	0.122	49.8
RandomScenario_seed-67	CRLI	**0.800**	**0.675**	**0.350**	**0.321**	7.4
	VaDER	0.700	0.573	0.146	0.126	3.1
RandomScenario_seed-79	CRLI	**0.733**	**0.608**	**0.215**	**0.184**	5.0
	VaDER	0.560	0.506	0.011	0.011	3.8
RandomScenario_seed-81	CRLI	**0.667**	**0.717**	**0.430**	**0.603**	0.2
	VaDER	0.549	0.546	0.166	0.326	5.7
RandomScenario_seed-88	CRLI	**0.212**	**0.907**	**0.074**	**0.273**	76.6
	VaDER	0.152	0.872	0.026	0.167	46.3
RandomScenario_seed-91	CRLI	0.416	0.968	0.275	0.642	250.6
	VaDER	**0.772**	**0.987**	**0.706**	**0.904**	183.3
RandomScenario_seed-94	CRLI	**0.363**	**0.946**	**0.211**	**0.535**	54.1
	VaDER	0.174	0.928	0.057	0.267	23.2
RandomScenario_seed-96	CRLI	**0.376**	**0.947**	**0.219**	**0.525**	75.8
	VaDER	0.236	0.927	0.106	0.409	44.7

Table 30. VaDER and CRLI performance on external CVI[cc]

[bb] While clustering RandomScenario_seed-53, we encountered a batch size error which caused VaDER to fail prematurely. For only that scenario, we changed VaDER batch size from 16 to 15.

[cc] For each experiment, the better performing method is bolded and underlined for each metric

Dataset	Purity_delta	RI_delta	ARI_delta	NMI_delta	Runtime_delta
RandomScenario_seed-15	0.118	0.052	0.136	0.12	153.225
RandomScenario_seed-23	0.2	0.101	-0.007	0.222	-0.59
RandomScenario_seed-26	0.222	0.222	0.572	0.393	1.222
RandomScenario_seed-30	-0.067	0	-0.047	-0.009	0.13
RandomScenario_seed-31	0.442	0.497	0.993	0.987	1.836
RandomScenario_seed-32	0.153	0.022	0.157	0.211	26.075
RandomScenario_seed-4	0.17	0.027	0.105	0.124	52.155
RandomScenario_seed-44	0.223	0.203	0.435	0.456	3.106
RandomScenario_seed-48	0.088	-0.002	0.042	0.143	2.377
RandomScenario_seed-53	0.138	0.016	0.088	0.117	5.321
RandomScenario_seed-59	0.133	0.047	0.033	0.139	3.219
RandomScenario_seed-6	0.266	0.102	0.366	0.423	1.181
RandomScenario_seed-62	0.34	0.043	0.271	0.458	30.564
RandomScenario_seed-67	0.1	0.102	0.204	0.195	4.288
RandomScenario_seed-79	0.173	0.102	0.204	0.173	1.175
RandomScenario_seed-81	0.118	0.171	0.264	0.277	-5.569
RandomScenario_seed-88	0.06	0.035	0.048	0.106	30.235
RandomScenario_seed-91	-0.356	-0.019	-0.431	-0.262	67.304
RandomScenario_seed-94	0.189	0.018	0.154	0.268	30.848
RandomScenario_seed-96	0.14	0.02	0.113	0.116	31.087

Table 31. Performance deltas (CRLI – VaDER)

5.5.2. Example #1: RandomScenario_seed-31

We remind the reader of RandomScenario_seed-31 properties and clustering result evaluation performance in Table 32. Figure 63 shows the full longitudinal trajectories for this dataset (train + test sets). This dataset was characterized by small sample size (n=280) and low inter-cluster separability on two variables (0 and 5). Despite that, CRLI was perfect across all four external CVI on RandomScenario_seed-31, greatly outperforming VaDER. Figure 64 shows the poor, non-discriminative latent representation learned by VaDER. None of the 2D visualizations demonstrate that VaDER was able to separate the samples into distinct clusters. The plots colored by ground truth (first column, orange and blue) show members from both clusters are interspersed amongst each other. Conversely, Figure 65 shows the highly discriminative latent representation learned by CRLI. The two clusters are very well separated and the cluster members themselves are predicted perfectly (second column) compared to ground truth, as reflected by the external CVI performance.

			CRLI	VaDER	Δ (CRLI - VaDER)
Properties	**timepoints**	15			
	vars	5			
	clusters	2			
	samples per cluster	200			
	train set size	280			
	test set size	120			
	noise	10			
	missingness	0.4			
	variable styles	0, 1, 2, 3, 5			
Performance	**Purity**		**1**	0.558	0.442
	RI		**1**	0.503	0.497
	ARI		**1**	0.007	0.993
	NMI		**1**	0.013	0.987
	Runtime		7.511	5.675	1.836

Table 32. RandomScenario_seed-31 properties and performance

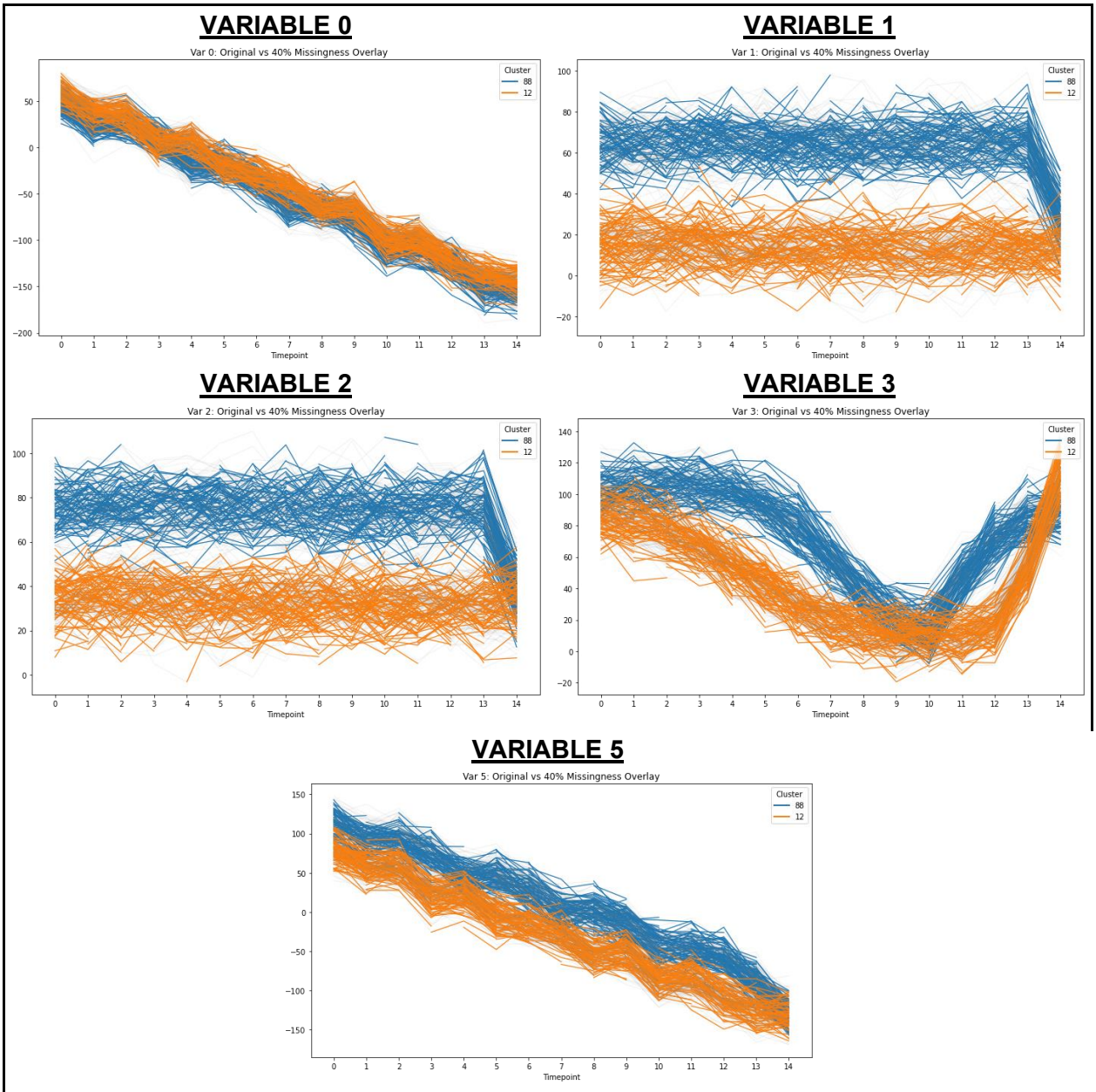


Figure 63. RandomScenario_seed-31 variable trajectories[dd]

[dd] As discussed in Methods, clusters in these figures are labeled by the cluster seed that was used to select the component parameter values underlying the time series variables.

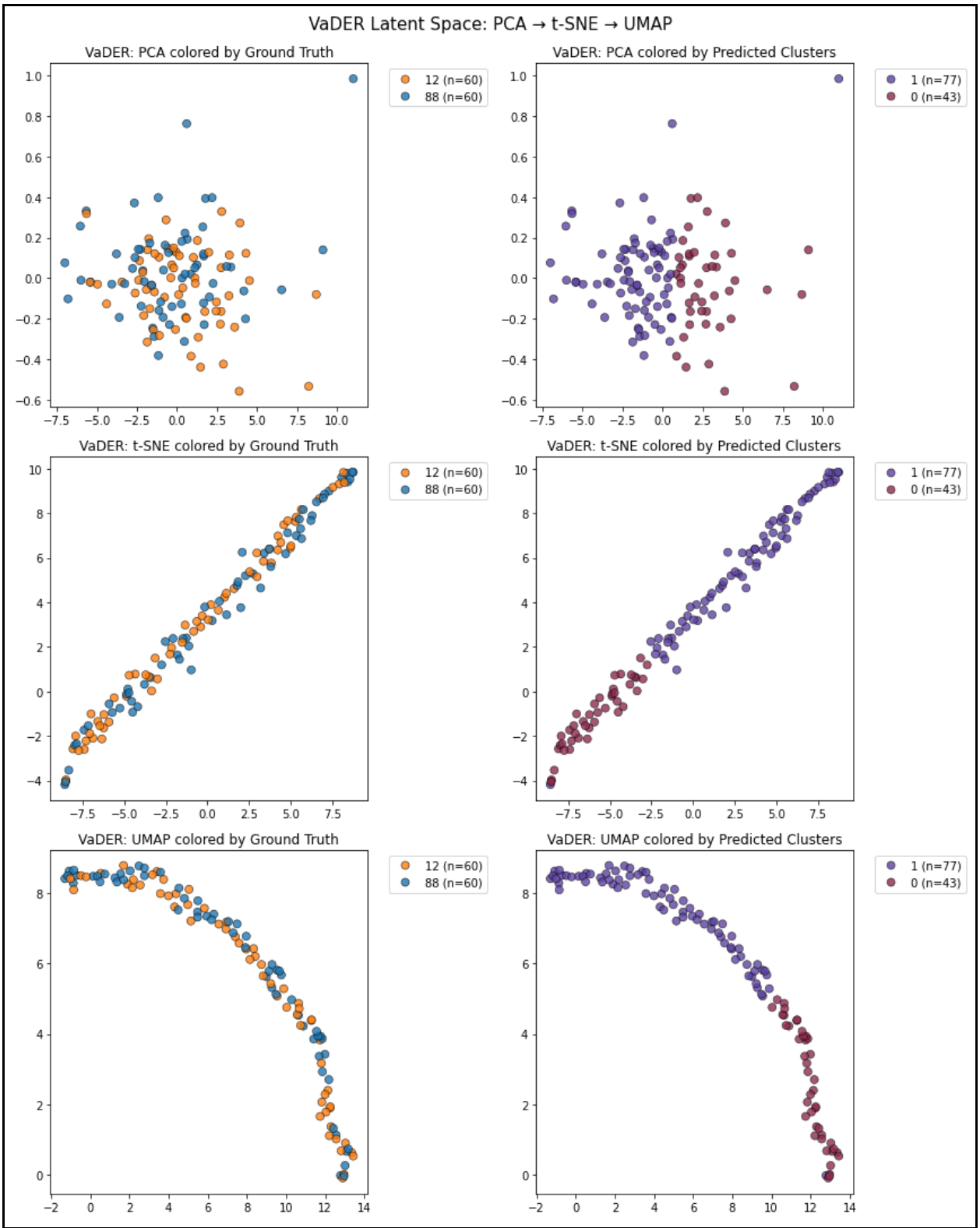


Figure 64. RandomScenario_seed-31 2D VaDER latent representation

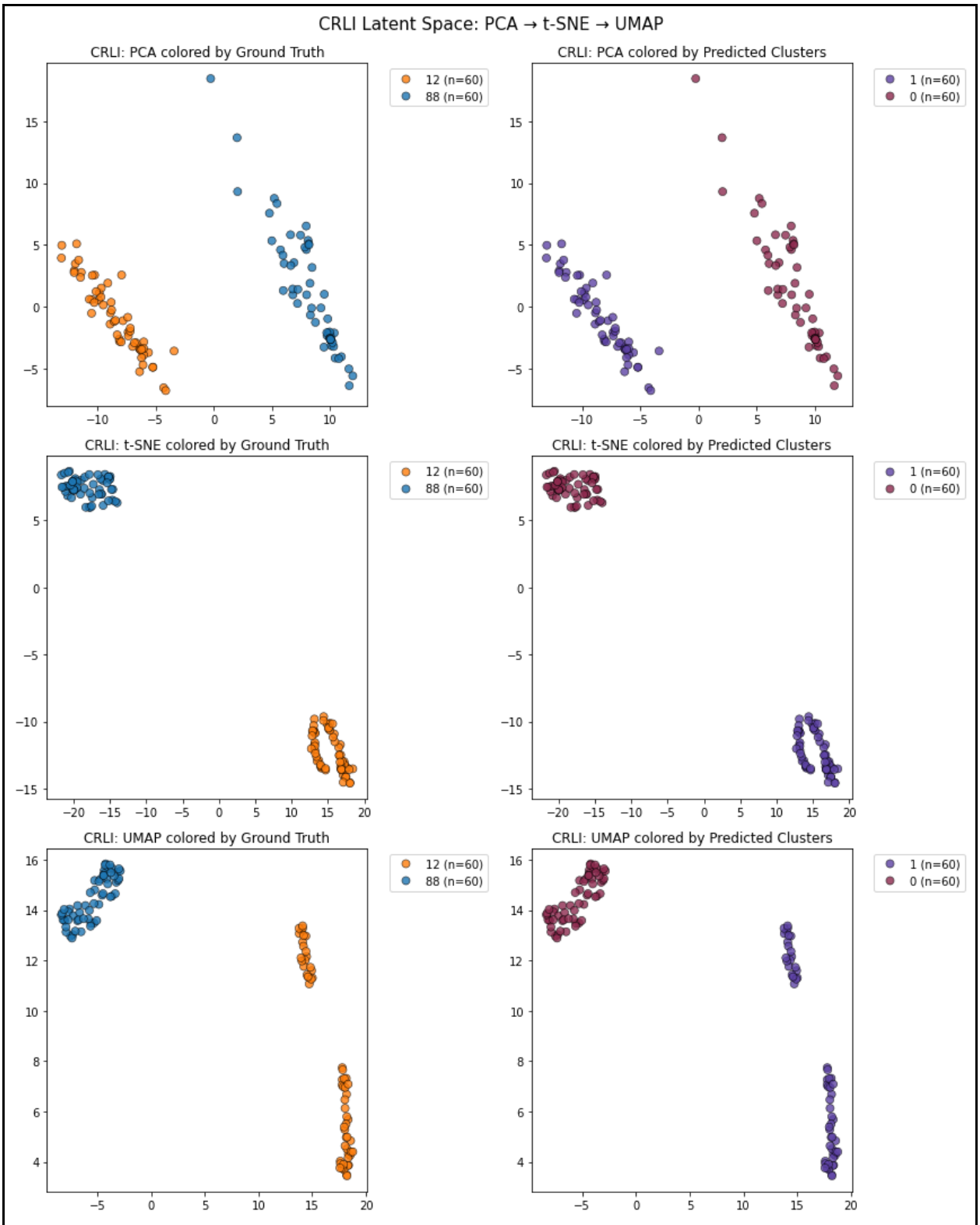


Figure 65. RandomScenario_seed-31 2D CRLI latent representation

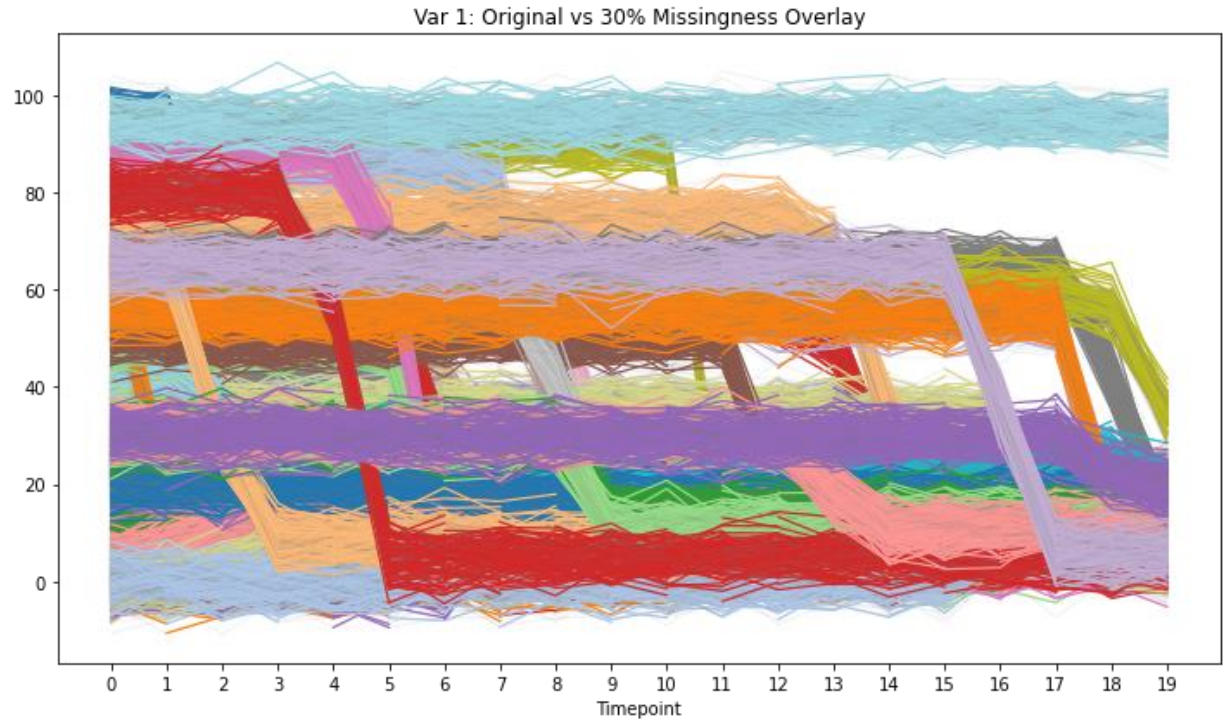
5.5.3. Example #2: RandomScenario_seed-91

We remind the reader of RandomScenario_seed-91 properties and clustering result evaluation performance in Table 33. Figure 66 shows the full longitudinal trajectories for this dataset (train + test sets). Notably, this dataset had the largest train set (n=7,000) and the most clusters (50) out of the 20 datasets we generated. It was the only dataset on which VaDER nontrivially outperformed CRLI, especially on the more conservative ARI. VaDER performing well when more training data is available is in line with Ma et al.'s findings.[33] In the latent representation learned by VaDER mapped to 2D (Figure 67), many tight, well-separated clusters are visible. However, not all cluster membership is predicted with 100% accuracy, which dampened ARI performance. For example, the red cluster at approximate coordinates (5,0) in the VaDER ground truth UMAP was split into two in the VaDER–predicted clusters plot. Figure 68 shows the latent representation learned by CRLI, which has far fewer well-separated clusters. The few samples that are separated into tight clusters are not consistently predicted to be in the same clusters. Due to high cluster number (50), legends were omitted for these 2D latent representation figures.

			CRLI	VaDER	Δ (CRLI - VaDER)
Properties	**timepoints**	20			
	vars	2			
	clusters	50			
	samples per cluster	200			
	train set size	7000			
	test set size	3000			
	noise	3			
	missingness	0.3			
	variable styles	1, 5			
Performance	**Purity**		0.416	**0.772**	-0.356
	RI		0.968	**0.987**	-0.019
	ARI		0.275	**0.706**	-0.431
	NMI		0.642	**0.904**	-0.262
	Runtime		250.579	183.275	67.304

Table 33. RandomScenario_seed-91 properties and performance

VARIABLE 1



VARIABLE 5

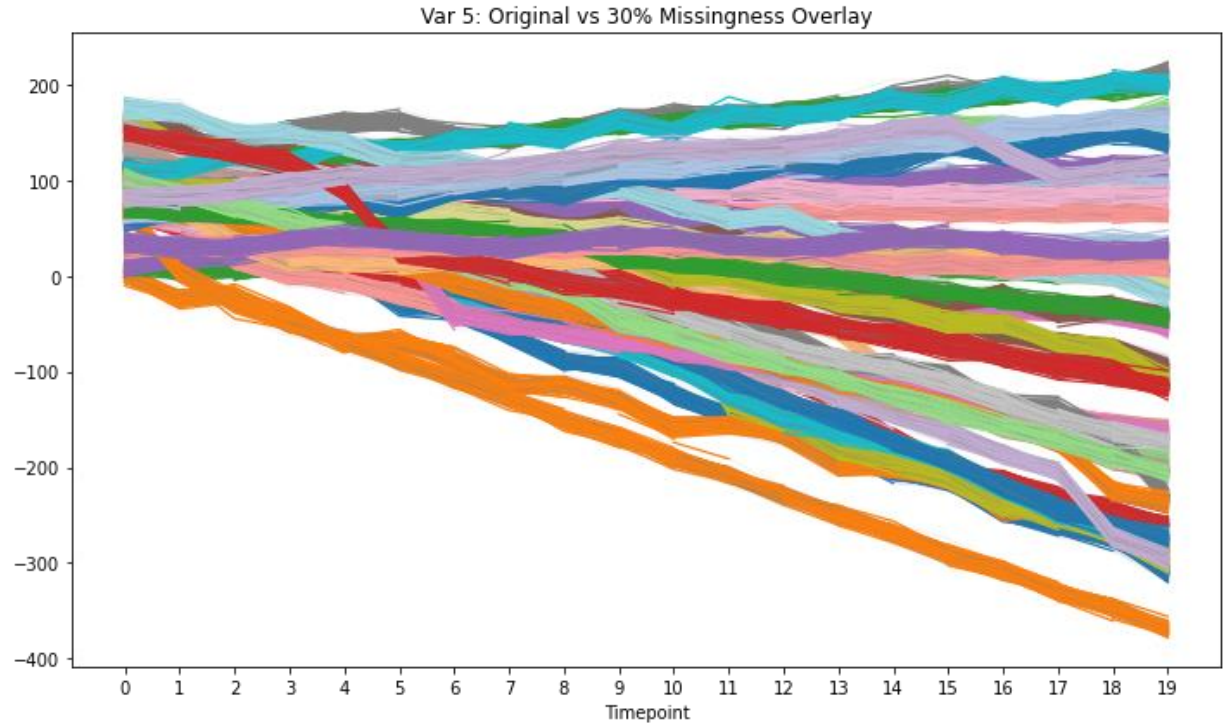


Figure 66. RandomScenario_seed-91 variable trajectories

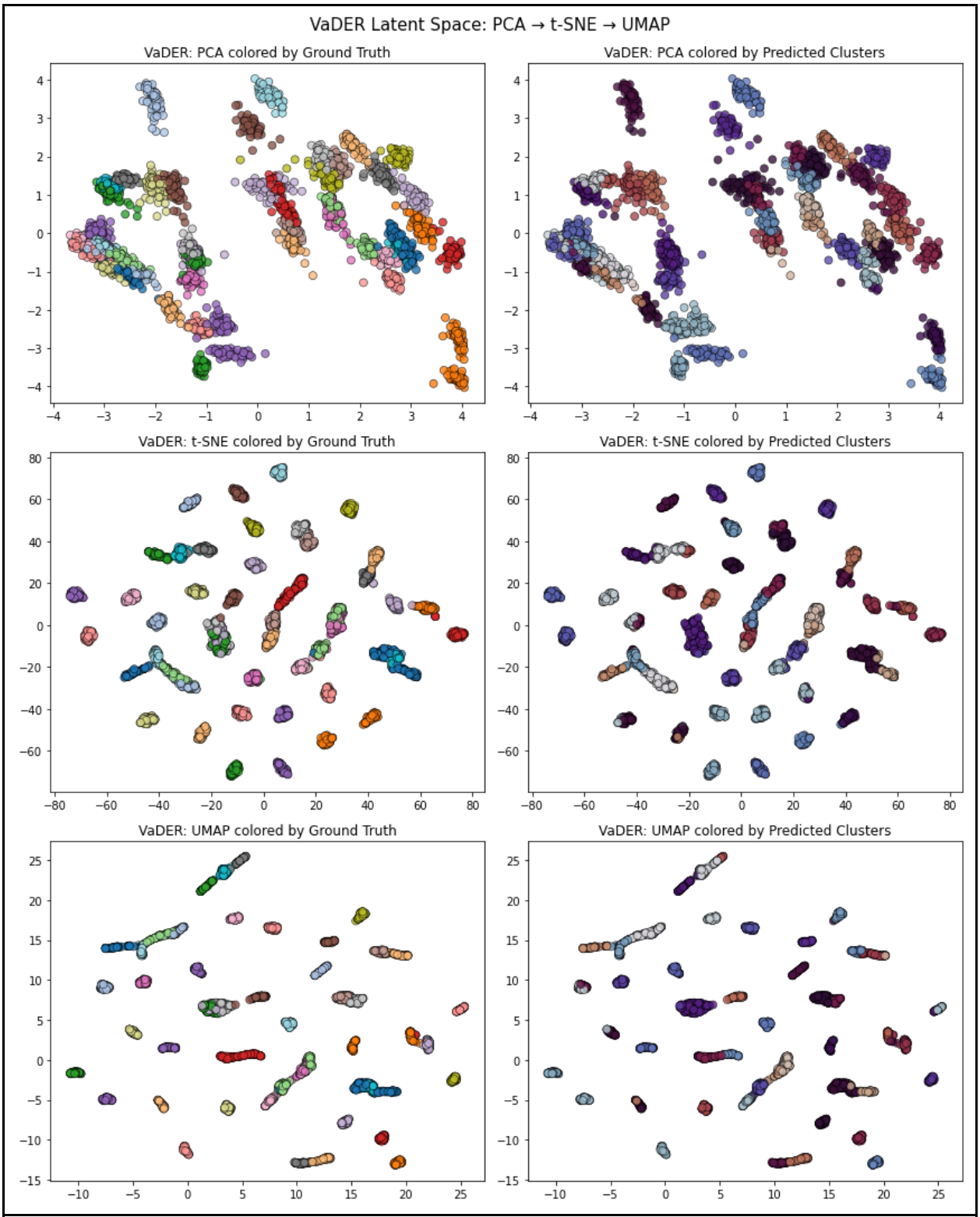


Figure 67. RandomScenario_seed-91 2D VaDER latent representation

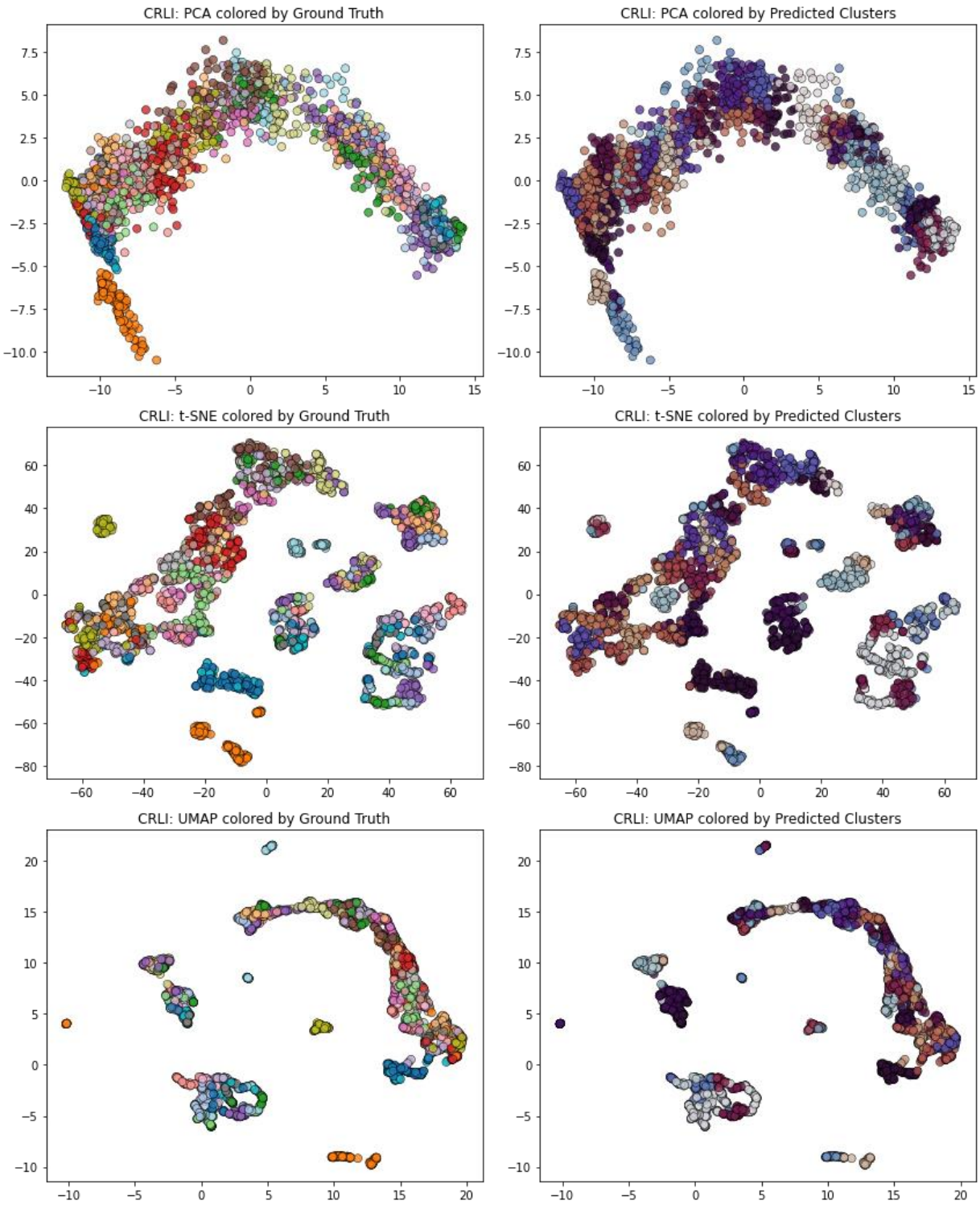


Figure 68. RandomScenario_seed-91 2D CRLI latent representation

5.5.4. Example #3: RandomScenario_seed-81

We remind the reader of RandomScenario_seed-81 properties and clustering result evaluation performance in Table 34. Figure 69 shows the full longitudinal trajectories for this dataset (train + test sets). While this dataset had no missingness, it was composed of very short time series (3 timepoints), meaning less total temporal data from the methods to learn from. On RandomScenario_seed-81, CRLI outperformed VaDER, but the latent representation learned by VaDER is visually more well-separated than the final cluster assignment would imply. Figure 70 shows the latent representation learned by VaDER, in which the cluster assignment (right column) does not reflect the separation very well. This is especially clear for ground truth cluster 49 (green, left column). Despite the proximity of the samples in the latent space, they are assigned to different clusters in the prediction. Furthermore, the predicted cluster sizes are quite imbalanced (351, 50, 49). Figure 71 shows the latent representation learned by CRLI, which suffered from a similar problem, though less severe. As with VaDER, the predicted cluster sizes were imbalanced (294, 82, 74). Both methods seemed to learn a representation that split samples into more than 3 clusters visually. Cluster 9 (orange) seemed to be well separated by both methods compared to the other two clusters. This could be a reflection of the longitudinal separability visible in Variables 0 and 6 in Figure 69.

			CRLI	VaDER	Δ (CRLI - VaDER)
Properties	**timepoints**	3			
	vars	5			
	clusters	3			
	samples per cluster	500			
	train set size	1050			
	test set size	450			
	noise	5			
	missingness	0			

	variable styles	0, 1, 3, 4, 6			
Performance	**Purity**		**0.667**	0.549	0.118
	RI		**0.717**	0.546	0.171
	ARI		**0.43**	0.166	0.264
	NMI		**0.603**	0.326	0.277
	Runtime		0.174	5.743	-5.569

Table 34. RandomScenario_seed-81 properties and performance

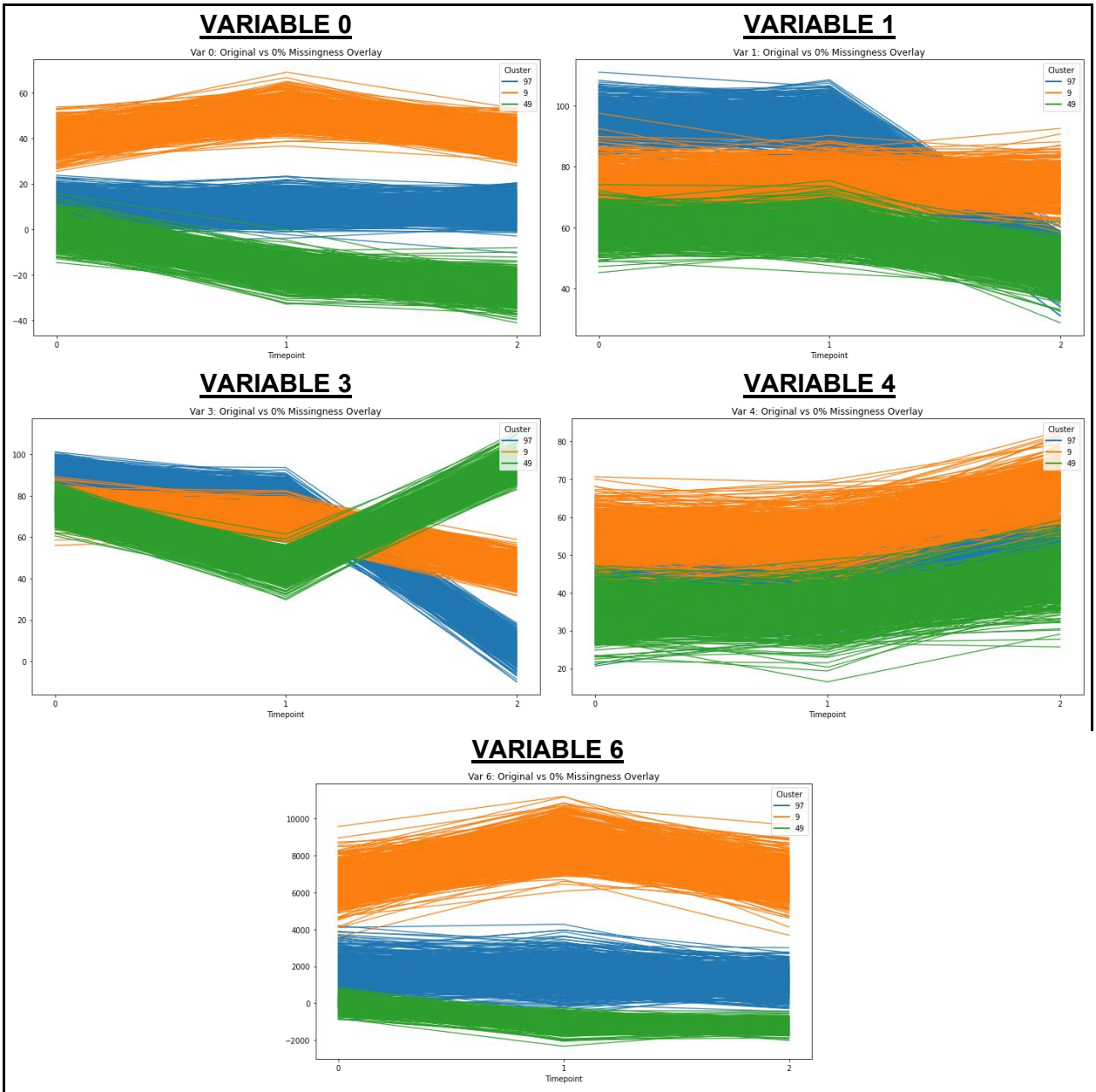


Figure 69. RandomScenario_seed-81 variable trajectories

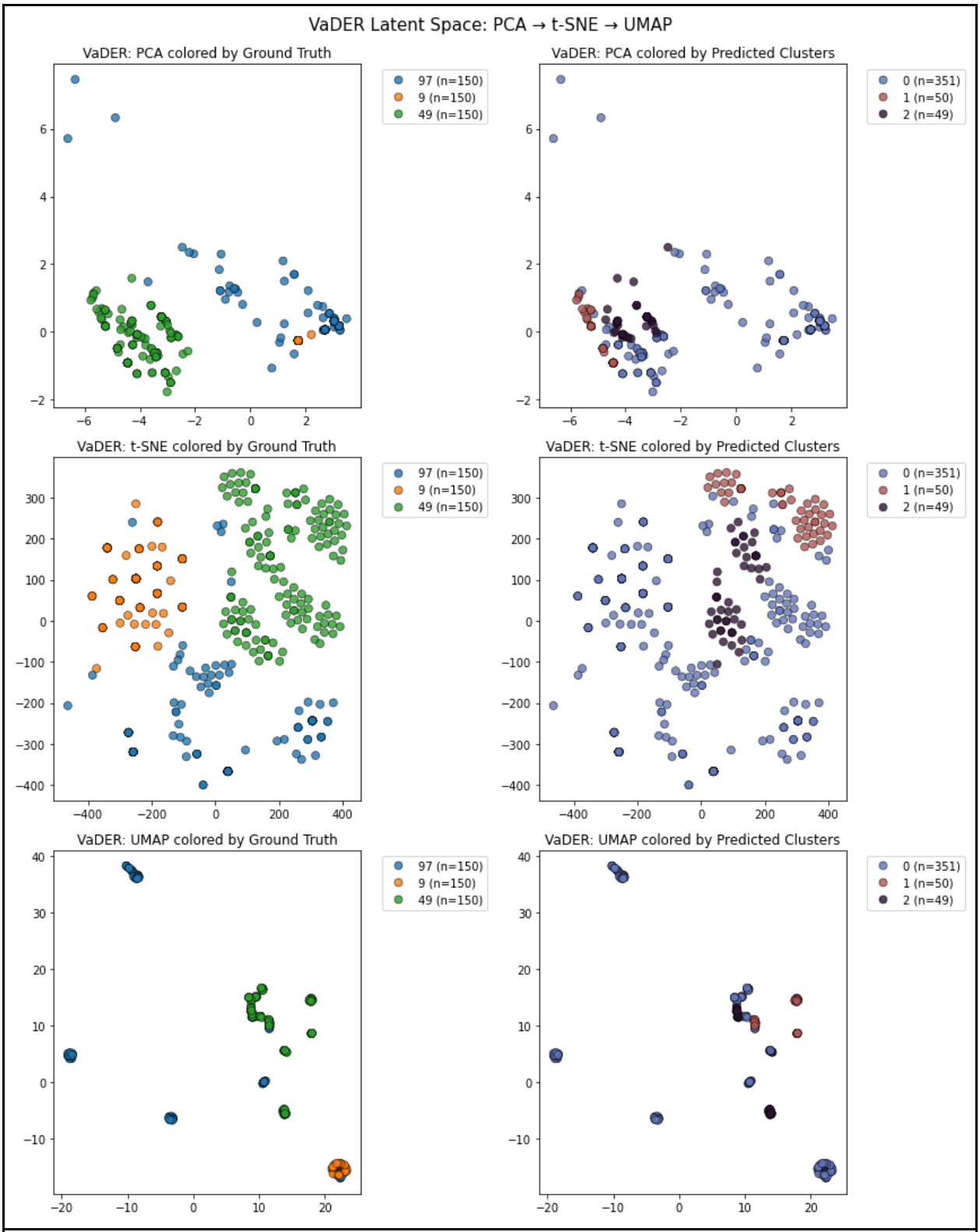


Figure 70. RandomScenario_seed-81 2D VaDER latent representation

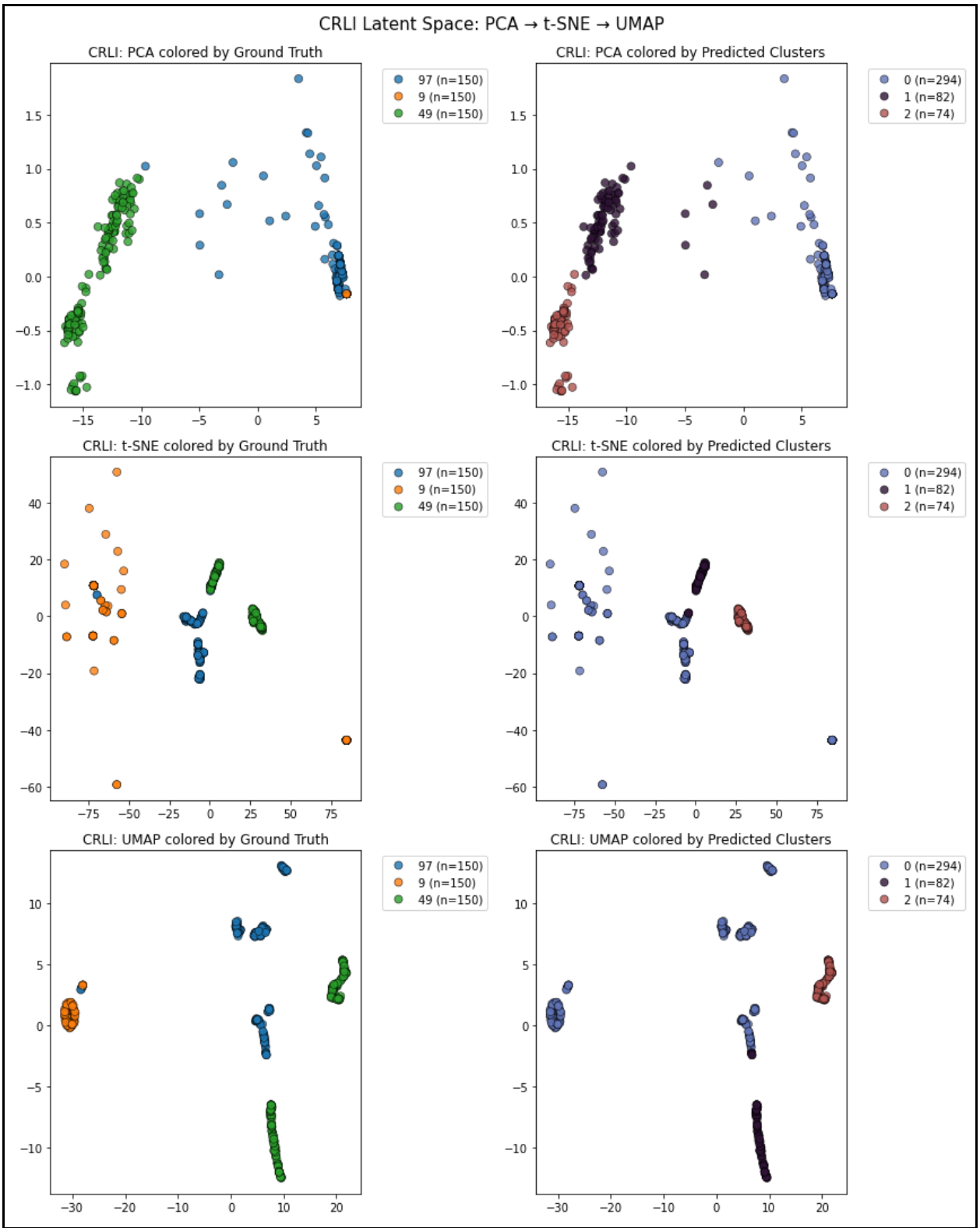


Figure 71. RandomScenario_seed-81 2D VaDER latent representation

5.5.5. Example #4: RandomScenario_seed-44

We remind the reader of RandomScenario_seed-44 properties and clustering result evaluation performance in Table 35. RandomScenario_seed-44 demonstrated the interplay between high longitudinal separability, low noise, and high missingness. Unlike in the previous examples, this dataset's 3 clusters rarely overlapped for any of the 5 variables (Figure 72). The latent representation learned by VaDER (Figure 73) was not impressive, showing poor separation between samples overall. However, cluster 14 (orange) was better separated than the others in the UMAP plot. CRLI isolated this cluster perfectly (Figure 74), but struggled to separate clusters 34 and 29 from each other. The ability of both methods to better distinguish cluster 14 may have been driven by its unique trajectories across Variables 5 and 6 (compared to the other two clusters which had similar trends or overlapping portions).

			CRLI	VaDER	Δ (CRLI - VaDER)
Properties	**timepoints**	20			
	vars	5			
	clusters	3			
	samples per cluster	50			
	train set size	105			
	test set size	45			
	noise	2			
	missingness	0.8			
	variable styles	2, 3, 4, 5, 6			
Performance	**Purity**		**0.867**	0.644	0.223
	RI		**0.855**	0.652	0.203
	ARI		**0.675**	0.24	0.435
	NMI		**0.765**	0.309	0.456
	Runtime		7.623	4.517	3.106

Table 35. RandomScenario_seed-44 properties and performance

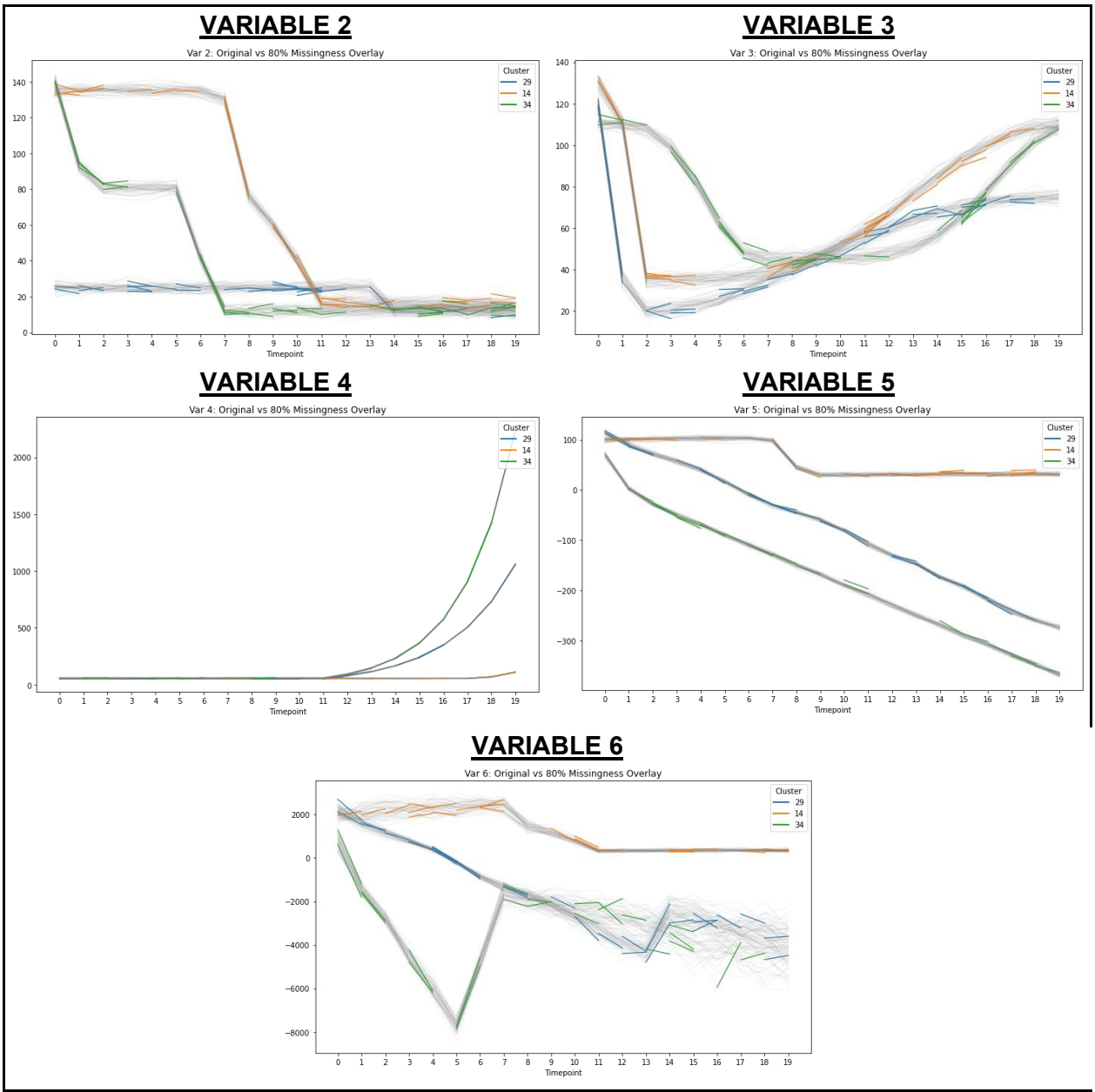


Figure 72. RandomScenario_seed-44 variable trajectories

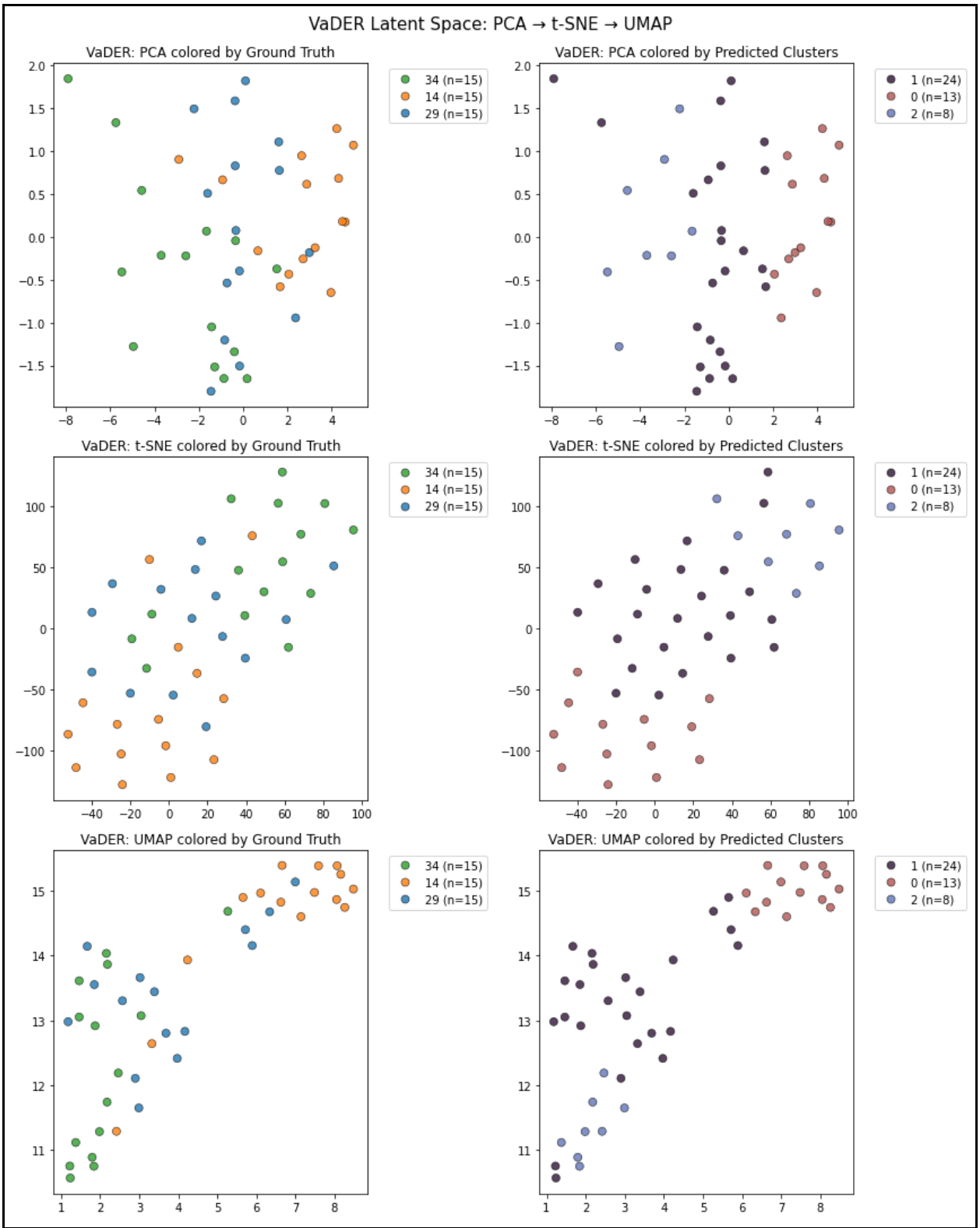


Figure 73. RandomScenario_seed-44 2D VaDER latent representation

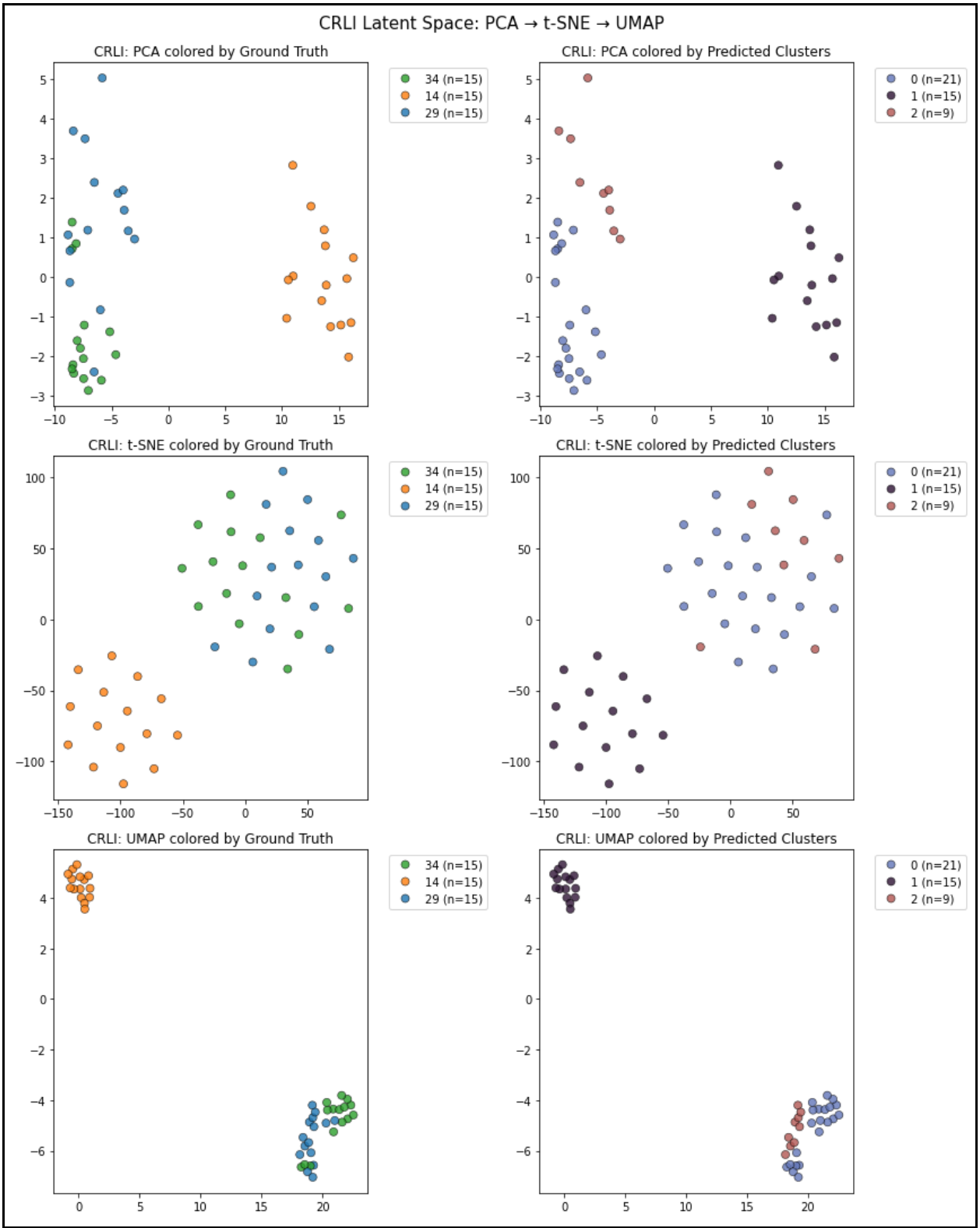


Figure 74. RandomScenario_seed-44 2D CRLI latent representation

5.5.6. Example #5: RandomScenario_seed-48

We remind the reader of RandomScenario_seed-48 properties and clustering result evaluation performance in Table 36. Compared to the previous example, RandomScenario_seed-48 had low longitudinal separability (Figure 75), due in part to very few timepoints. Compared to RandomScenario_seed-81 (Section 5.5.4), which also only had 3 timepoints, this dataset had more clusters (20) and a higher missingness proportion (0.4). Both VaDER and CRLI struggled to perform well, as seen in Figure 76 and Figure 77, respectively. The UMAP plots of both methods' latent spaces show that none of the learned separation corresponds to ground truth label distinguishability. While this is not reflected in the inflated RI scores, it is captured by the poor ARI performance.

			CRLI	VaDER	Δ (CRLI - VaDER)
Properties	**timepoints**	3			
	vars	6			
	clusters	20			
	samples per cluster	100			
	train set size	1400			
	test set size	600			
	noise	4			
	missingness	0.4			
	variable styles	0, 1, 2, 3, 4, 6			
Performance	**Purity**		**0.243**	0.155	0.088
	RI		0.875	**0.877**	-0.002
	ARI		**0.059**	0.017	0.042
	NMI		**0.294**	0.151	0.143
	Runtime		10.15	7.773	2.377

Table 36. RandomScenario_seed-48 properties and performance

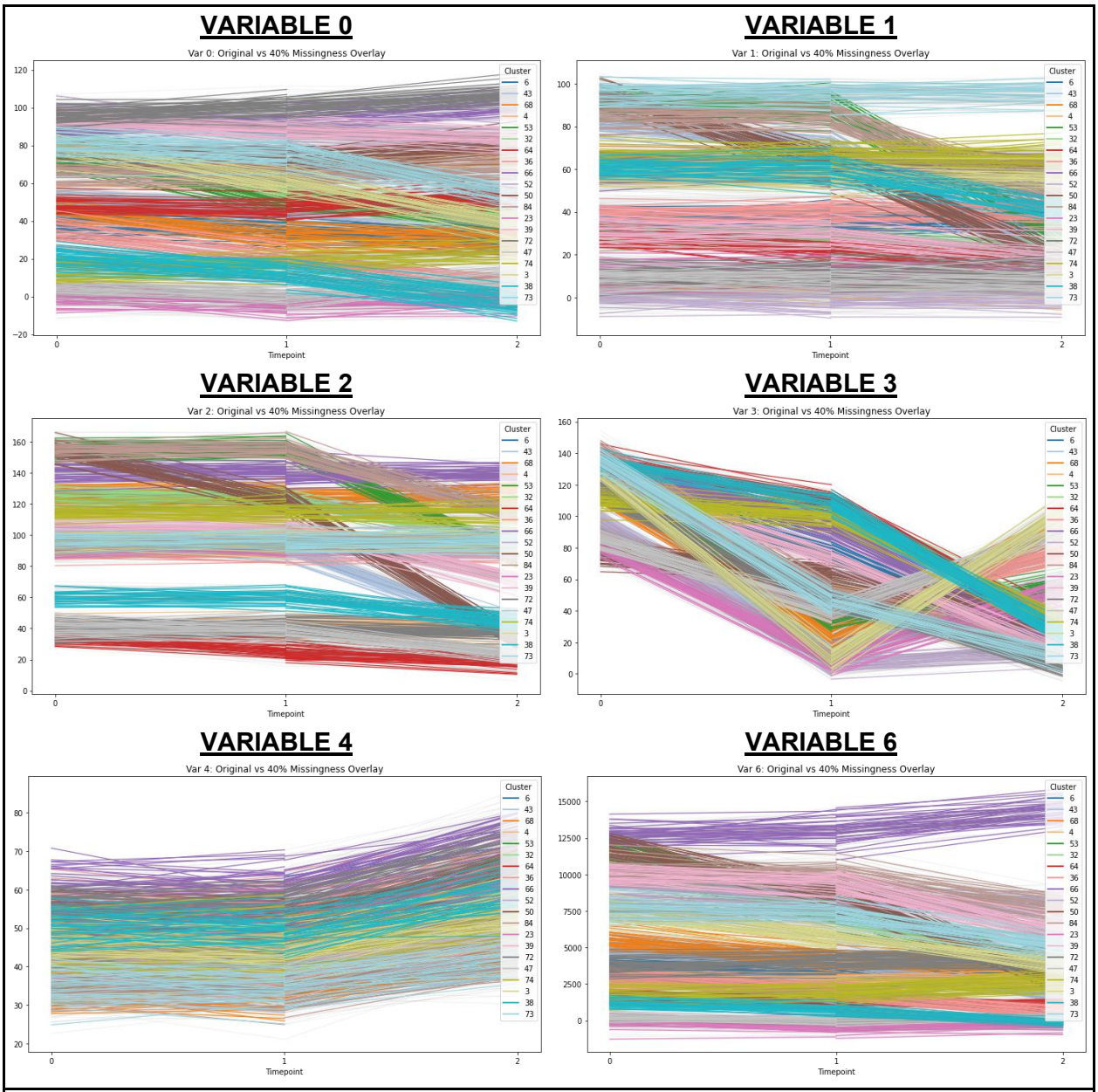


Figure 75. RandomScenario_seed-48 variable trajectories

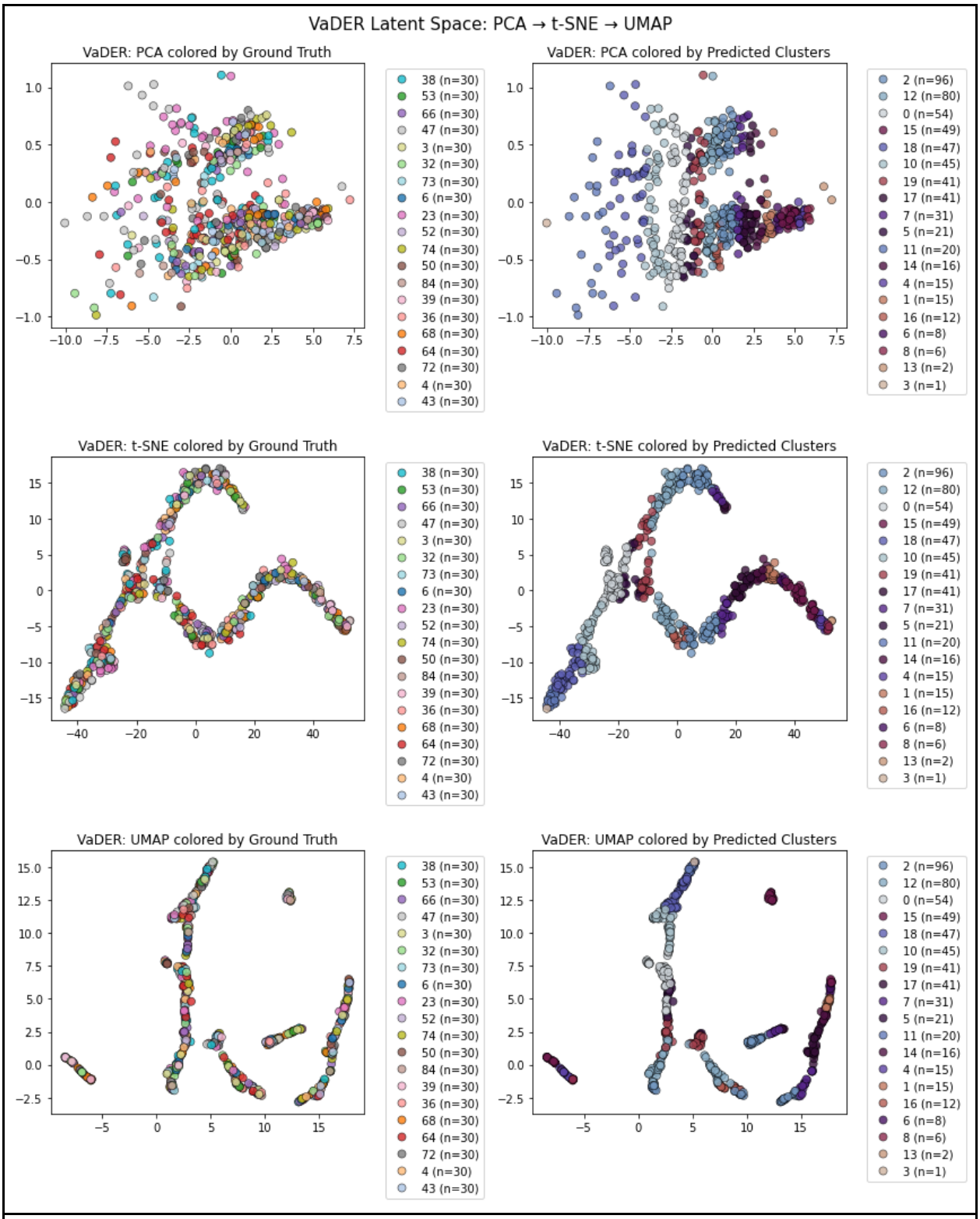


Figure 76. RandomScenario_seed-48 2D VaDER latent representation

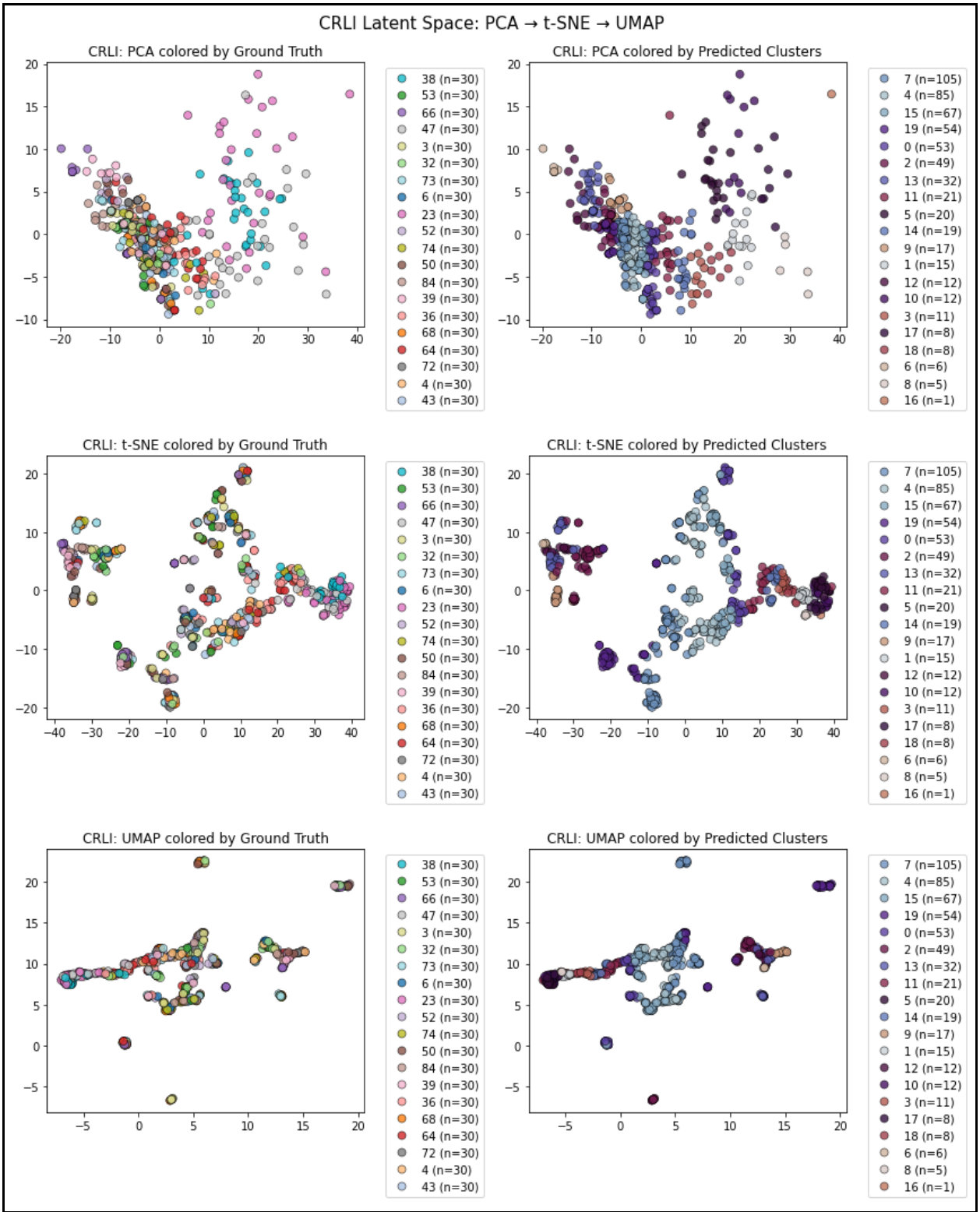


Figure 77. RandomScenario_seed-48 2D CRLI latent representation

5.5.7. Experimental takeaways

In summary, we (1) reinforce previous work which has recommended ARI as the standard for external clustering validation, (2) show CRLI's outperformance of VaDER under most data conditions tested, and (3) establish dataset properties that can potentially impact performance of one or both methods. ARI crossed 0.5 in only 4/40 cases (compared to RI, which crossed 0.8 in 28/40 cases). While CRLI outperformed VaDER on 17/20 datasets for ARI, the magnitude of the difference between the two methods was less than 0.2 for the majority of datasets. In our analyses of individual experiments, we found that both methods benefited from longer time series and lower missingness. CRLI was able to perform better in trickier conditions (high missingness, few samples) but was outperformed by VaDER when many training points were available.

5.6. Discussion

5.6.1. Summary

In this Aim, we demonstrated the ability to generate any number of synthetic multivariate time series (MVTS) datasets, without the need for existing data, and with fine-grained control over important time series properties, in a way that can mirror commonly observed biomedical patterns (Section 5.4.1). We utilized the Python package mockseries to achieve this and present the approach as a complement to MVTS clustering method benchmarking which typically has relied on real-world datasets that are often (1) not multivariate, (2) not from the medical domain, and (3) not having much missing data. We generated 20 MVTS datasets with varying lengths, missingness proportions, noise levels, cluster numbers, and sample sizes (Table 26) and used them to evaluate the performance of VaDER and CRLI, two state-of-the-art incomplete MVTS deep clustering methods.

As expected, CRLI, developed after VaDER and with VaDER's shortcomings in mind, outperformed VaDER in the majority (18/20) of scenarios (Table 31). However, in the case of RandomScenario_seed-91 (Section 5.5.3), VaDER was superior, likely due to the low missingness (30%) and large amount of training data available (n=7,000), which is in line with Ma et al.'s critique of VaDER that it performs well only on large datasets.[33] Optimal method performance on a given dataset is the result of achieving a delicate balance between the reconstruction and clustering tasks, which is not trivial. In some cases, the latent representation learned by VaDER was visually impressive, though ultimately the cluster assignment was unsatisfactory. This warrants further investigation, as discussed in Future Directions. We also showed that, for all but two cases (Table 31: RandomScenario_seed-23 & RandomScenario_seed-81), CRLI took longer to run, though the performance gains likely make this worth it.

Our work underscores the importance of reporting multiple, complementary cluster validation indices (CVI). In their VaDER benchmarking experiments, de Jong et al. reported only cluster purity. When comparing CRLI performance to VaDER's, Ma et al. reported only Rand Index (RI) in the main text and Normalized Mutual Information (NMI), purity, and accuracy in the supplement. Importantly, we report performance on Adjusted RI (ARI), which, as discussed earlier (Section 2.4.4.2), is a more conservative measure of performance than others. This is because the ARI calculation is (i) independent from the number of clusters and (ii) adjusts for chance.[88,89] Our trajectory visualizations, which are often absent from benchmark studies, demonstrate the need for a clusterability, or longitudinal overlap, measure, which quantifies how discernable the clusters are from each other per variable. This concept is obvious upon visual inspection of trajectories but is not reflected in the dataset properties, like number of clusters, time series length, and number of variables, that are typically reported, as we did in this Aim.

5.6.2. Limitations

5.6.2.1. Properties not explored

There are some properties that we did not explore in more depth and others that were out of scope entirely. For example, our search space did not include very long time series, like those that could result from telemetry monitoring. Our simulations were inspired by longitudinal EHR and observational cohort datasets and as such did not address areas of biomedicine that have a higher sampling rate. Another limiting factor of our workflow was that variable style design was manually curated. Thus, though we generated 7 unique variable styles, there are many other possible patterns of biological or clinical relevance that may be worth modeling. Within each of our datasets, clusters were of the same size. This is not a totally faithful representation of real patient cohorts, where distinct patterns may be disproportionately spread throughout a population. We also only modeled one missingness mechanism, missing completely at random (MCAR),

though others, like missing at random (MAR) and missing not at random (MNAR) are prevalent in biomedicine.[56,57] Finally, for the properties we did explore, and their corresponding search spaces we defined, we limited our data generation process to 20 scenarios due to time and computational constraints. Though this did not nearly cover the entire property search space, which is technically infinite, it gave us a solid foundation on which to build future benchmarks. Since this Aim only considered with time series generation, we neglected to generate and perform analyses of static features and how they can vary (enrichment vs. depletion) across longitudinal clusters. This includes components of fair unsupervised learning approaches.[264]

5.6.2.2. Normalization/standardization were not applied

We did not apply any normalization or standardization techniques to the raw time series samples. For reference, in the original VaDER publication, benchmark dataset time series were standardized to zero mean and unit variance, while all ADNI and PPMI (longitudinal observational studies)[ee] timepoints were normalized relative to baseline.[54] Our intention was to assess how well methods dealt with MVTS data "out of the box", but a more robust analysis could have included comparisons with preprocessed datasets.

5.6.2.3. Cluster difficulty quantification was not performed

In the unlikely event that two datasets shared all the same dataset properties, they could have had different variable style properties due to our random cluster seed selection and mockseries parameter generation process. This was by design to explore complexity across different trajectory shapes. However, this meant that scenarios had varying, unquantified amounts of longitudinal overlap between clusters for one or more variables. De Jong et al. quantified this concept in their vector autoregressive (VAR) process simulations with clusterability parameter $\boldsymbol{\lambda}$,

[ee] ADNI = Alzheimer's Disease Neuroimaging Initiative; PPMI = Parkinson's Progression Markers Initiative

where a higher value corresponds to easier separability between simulated clusters.[54] Without an analogous metric in our analysis, we were unable to broadly assess the effect that clusterability had on method performance, except on a case-by-case basis when we visually inspected full variable trajectories.

5.6.2.4. Additional evaluation metrics could be explored beyond the four reported

Besides the four we reported, other external CVI described in the literature, such as F-measure, Entropy, Adjusted Mutual Information (AMI), and Clustering Accuracy, may have been additional valuable comparison points.[27,28] Additionally, though for this aim we had ground truth labels against which to evaluate performance (using external CVI), in practice, when true cluster membership is unknown, internal CVI are used to determine the optimal cluster number for a dataset, as discussed in Section 3.4.5. Running VaDER and CRLI on multiple *K* values for each dataset, as done in Aims 1 and 2, and calculating internal CVI would have allowed us to see if the internal CVI-informed optimal cluster number differed from the ground truth cluster number. Lastly, since these methods are nondeterministic, we could have run each experiment multiple times and reported average CVI results and standard deviations as measures of robustness and stability.[33] These additional analyses were not performed due to time and computational constraints.

5.6.3. Future directions

The work performed in this aim sparked a number of interesting additional research directions. In order to better understand what is driving performance and runtime differences, we could model per-scenario performance as a function of dataset properties and use methods like SHAP to quantify feature importance.[265] We could also devise a clusterability metric, perhaps by drawing from the time series complexity literature, to include in such a modeling analysis that would capture the inherent differences between cluster seeds that are not reflected in the dataset

properties we reported.[266] For a more targeted analysis of the influence of specific dataset properties, we could take a subset of promising scenarios and modify a single property at a time while holding all others constant.

To isolate and evaluate only the quality of the representation learned by either method, as opposed to the final cluster label assignment, we could take the 2D representations mapped by PCA, t-SNE, or UMAP and apply one standard clustering method (like k-means) to both VaDER and CRLI outputs. We could then compare performance with the full end-to-end methods to see what, if any, performance impact the representation learning step has versus the cluster assignment step. Along these lines, McConville et al. competed with and, in some cases, outperformed state-of-the-art deep clustering methods by replacing the clustering network with a framework that (1) applies UMAP to the encoded representation and (2) clusters this new embedding with a shallow clustering algorithm (Gaussian Mixture Model).[263] In parallel, we could compute per-cluster measures of compactness and/or dispersion, purity, and separation to inform individual cluster "confidence" values. These would complement the global external CVI, which is computed on all test samples, with a more specific quantification that could encourage certainty in a subset of clusters, even if overall performance is unsatisfying.

5.7. Conclusion

In Aim 3, we assessed the ability of CRLI and VaDER to detect meaningful trajectories in synthetic datasets under diverse data constraints. In the process, we laid the foundation for a simulation-based approach to benchmarking IMVTS clustering methods in general. Previous comparable benchmarks have (1) been limited in their scope of explored properties, (2) used real-world dataset archives originally designed to assess supervised classification method performance, and/or (3) not been attentive to missingness beyond trivial or default imputations.[28,33,54] Inspired by longitudinal patterns found in Aims 1 and 2, we used mockseries, a synthetic time series generation package written in Python, to design a number of time series variables analogous to common biomedical trajectories. Our data generation process (DGP) allows for rapid production of very diverse datasets by random sampling from variable building blocks and across a number of MVTS properties, like cluster number, time series length, noise, and missingness. This DGP enabled us to assess the ability of one-stage IMVTS deep clustering methods (VaDER, CRLI) to detect complex trajectories under diverse constraints. Unlike Aims 1 and 2, in this aim we were able to compare predicted clusters against ground truth clusters and report external clustering validation metrics (CVI), including purity, Rand Index, Adjusted Rand Index, and Normalized Mutual Information.

Overall, we reported that CRLI overperformed VaDER on these external CVI across the majority of datasets. However, we note that some metrics can be forgiving, like RI, and others can be overly conservative, like ARI. We also note that lower CVI did not always correspond to poor latent representation partitions, as observed in the 2D PCA, t-SNE, and UMAP visualizations we produced. A more comprehensive set of separation quality metrics that are calculated per-cluster would be a valuable addition to future work. CRLI's performance even in low sample size and high missingness contexts was impressive, but inclusion of more datasets with larger sample

sizes (n = 5,000+) would have provided additional confirmation, beyond the single dataset, that VaDER can compete when training data is abundant. Since we did not conduct hyperparameter tuning for either method, or multiple runs per dataset, we cannot rule out potential performance gains that would have resulted if we had. Since, in practice, these methods are deployed in the absence of ground truth labels, a comparative analysis between performance on internal and external CVI is warranted, and would make assessment alongside Aims 1 and 2 more feasible.

Our simulation-based approach in Aim 3 complements the real-world data-based exploration of IMVTS deep clustering in the earlier aims. In Aim 3, we developed a tunable and modular approach to IMVTS data generation and increased the quantity and variety of longitudinal data constraints under which VaDER and CRLI have been placed to date. By doing so, we were able to identify those scenarios where these methods may be most useful with regard to biomedical research data exploration. In particular, we demonstrated those data conditions under which each respective method can thrive and/or fail, which can inform future work on when and how to deploy these methods on real biomedical use cases.

6. Summary

In this dissertation, we sought to characterize how and when one-stage MVTS clustering methods are useful in biomedical research data exploration. These particular methods combine time series imputation, representation learning, and clustering (Figure 78) to group incomplete multivariate time series data into distinct clusters that are maximally dissimilar between clusters and minimally dissimilar within clusters. At the time of writing, VaDER (2019) and CRLI (2021) are the two state-of-the-art methods in this area. They are particularly suitable for clinical subgroup analyses

(disease progression, treatment response, developmental trajectories, patient monitoring) because longitudinal medical data is characterized by high missingness, multiple outcomes measures of consequence, and complex temporal relationships, all of which these methods are designed to address.

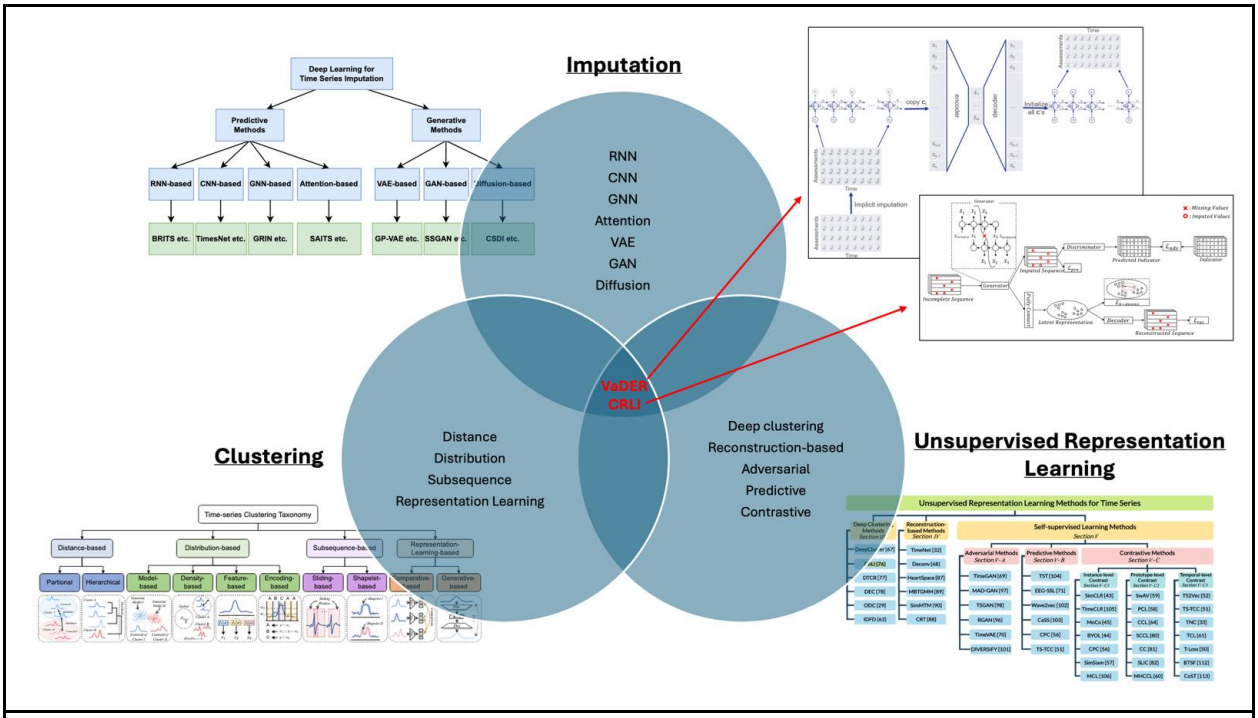


Figure 78. One-stage deep representation learning for MVTS clustering: VaDER and CRLI[27,33,54,69,81]

We selected 3 contexts (Figure 79, Table 37) to study the application of these recently developed methods to real-world longitudinal biomedical datasets. First, we assessed the ability of CRLI to detect meaningful trajectories in the EHR, characterized by sparse, irregularly sampled time series. Second, we assessed the ability of CRLI to detect meaningful trajectories in the ABCD Study, a prospectively designed longitudinal observational cohort, characterized by greater regularity of measurement, higher retention, and multimodality. Third, we assessed the ability of CRLI and VaDER to detect trajectories in synthetic datasets, characterized by systematic variation across a number of dataset properties of interest. In this section, we summarize each

Aim: our key findings, limitations, and future directions. We then discuss cross-cutting observations that emerged from a holistic review of all Aims.

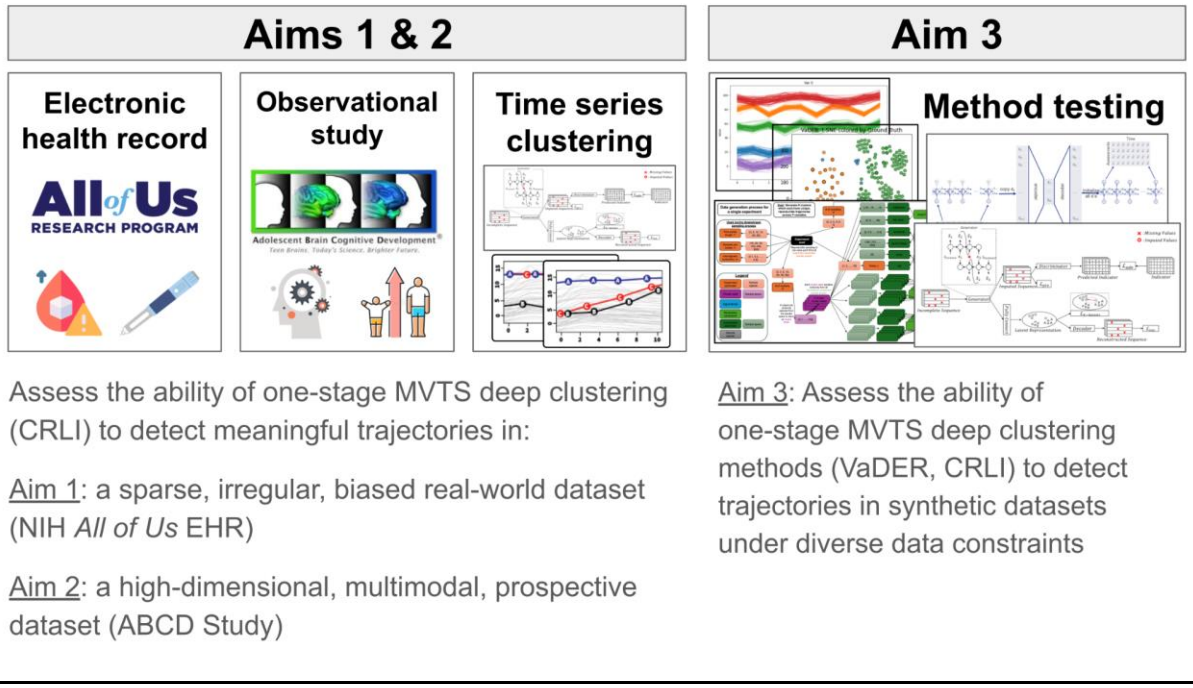


Figure 79. Graphical summary of dissertation aims

	Aim 1	**Aim 2**	**Aim 3**
Research area	Glucagon-like peptide-1 (GLP-1) receptor agonist (RA) treatment response	Physical & mental adolescent development	Benchmarking for MVTS clustering
Data source type	Electronic health record	Longitudinal observational cohort	Synthetic
Dataset	NIH *All of Us* Research Program	Adolescent Brain Cognitive Development (ABCD) Study	Newly-developed framework to generate random MVTS datasets
Tools	CRLI	CRLI	mockseries, VaDER, CRLI
MVTS clustering applicability	Multiple treatment outcome measures collected in routine clinical care	Many multidomain developmental measures assessed yearly	Range of dataset & time series properties, incl. # of variables, missingness rate, sample size, noise, time series length
Data challenges	Short time series lengths, irregular sampling rates, data sparsity, nonlinearity, variable inter-correlations		

Table 37. Aims overview

6.1. Aim 1

In Aim 1, we assessed the ability of CRLI to detect meaningful multivariate trajectories in a sparse, irregularly sampled, biased real-world data source, the electronic health record (EHR). In the age of big data, the EHR is an increasingly utilized source for biomedical research data exploration. The biomedical value of EHR data lies in its natural reflection of routine clinical care experienced by patients. However, temporal EHR data is marred by irregular measurement intervals, high missingness, and multiple biases (selection, measurement, time-related).[58] Our focus was to assess how CRLI handled these hurdles in the context of identifying GLP-1 medication (semaglutide, dulaglutide, etc.) treatment response subgroups in the NIH *All of Us* Research Study (v7 CDR).

6.1.1. Key findings

We selected a cohort of 336 *All of Us* participants taking GLP-1 medications for at least 2 years, who had measurements of HbA1c, serum creatinine, BMI, DBP, and SBP at least once in the 6 months prior to starting the medication and twice in the 2 years after starting the medication (Figure 32). Due to discordant internal CVI results (Figure 36), we reported a 5-cluster result and a 2-cluster result. Both clustering results mapped to clinically relevant ranges across multiple variables, showing that CRLI was able to identify quantitatively and qualitatively distinct clusters. In both clustering results, we saw HbA1c decrease for all clusters, as expected. Interestingly, BMI was quite different between clusters, but relatively stable within each cluster. We generated mean lineplots with 95% confidence intervals for each cluster, allowing us to visualize cluster separation and lability. In the context of the overarching question of how and when IMVTS deep clustering methods are useful in biomedical data exploration, the most important findings were that (1) CRLI can be used to identify post-treatment multivariate response trajectories in the EHR and (2) this

is possible despite a small cohort (n=336) and infrequent measurements. However, clinical interpretation is challenging due to high dimensionality, and single-timepoint (cross-sectional) clustering may be just as effective for identifying treatment response phenotypes, as evidenced by outcome measure stability post-treatment in the longitudinal clusters we identified.

6.1.2. Limitations

Due to the multiple repeat measurement requirements we imposed, our cohort selection bottleneck was severe: we retained only 336 (3.2%) of the 10,367 GLP-1 users in *All of Us*. This raises questions about the generalizability of our results to the broader GLP-1 user population and whether our clusters could be replicated in another comparable dataset. Furthermore, from a statistical perspective, our approach assumed missingness completely at random (MCAR) since the methods' imputation processes did not have access to any static data that could have explained missingness patterns in the longitudinal variables (Section 2.3.1.3) A lack of agreement between the 4 internal CVI we used reduced confidence in the "true" underlying cluster number. Future work could incorporate more recently developed CVI or formalize a strategy to pick a cluster number when there is disagreement. Lastly, we did not have reliable structured dosage data, did not have dispensing information, and cannot be certain of medication adherence over the 2-year medication use period studied. In the context of the overarching question of how and when IMVTS deep clustering methods are useful in biomedical data exploration, the implications of these limitations are that (1) trajectories identified in the selected cohort may not be generalizable to the population being studied, (2) a clustering experiment may need to be repeated many times and assessed by many metrics to ensure confidence in results, and (3) EHR data complexities, like excessive missingness or data quality issues, may be too difficult to overcome to select a suitable cohort for longitudinal clustering.

6.1.3. Future directions

Since completing this Aim, *All of Us* has released v8 (Feb 2025) which contains 20,793 GLP-1 users.[267] Revisiting our analysis with this updated data will surely increase our final cohort size, both because there are more participants overall taking GLP-1 medications, but also because some participants who would not have met our criteria due to insufficient time spent on the medication (< 2 years) will likely now qualify. This is important because it would increase the chances of generalizability to the GLP-1 user population and increase confidence in the longitudinal clusters identified.

We limited our analysis to 5 outcome measures due to (1) relevant T2D treatment monitoring guidelines and (2) clinical interpretability. However, patients may have other measurements (Complete Blood Count, Metabolic Panel) drawn at routine visits which could be incorporated into our analysis as additional longitudinal outcome measures. It may become difficult to interpret the clusters clinically as the number of variables grows, but recent advances in explainability (Section 6.5.2.3) could aid in understanding, along with visualizations like PCA, t-SNE, and UMAP, which were generated in our Aim 3 results. As biomedical datasets increase in breadth and depth (number of participants, number of variables, number of timepoints), it will be important to demonstrate that CRLI can identify meaningful trajectories across more than a handful of variables since big biomedical data exploration often emphasizes finding patterns in high-dimensional contexts.

To characterize our clusters beyond quantitative internal CVI values and reference range alignment of the longitudinal variables that were clustered, we could perform association analyses to describe the between-cluster differences in comorbidities, concurrent medication use, social determinants of health, and other participant characteristics.[2] This is important because detecting

“meaningful” trajectories is also a qualitative goal. In the context of biomedical data, clinical interpretability and actionability are as important, if not moreso, than performance on quantitative metrics.

Lastly, to assess the generalizability of our results, we propose validating these clusters in other, comparable real-world datasets, like the UK Biobank, and even clinical trials. If application of CRLI in these other contexts yields similar trajectory subgroups, we will have confidence that these trends are not unique to just *All of Us*. Though our assessment of CRLI ability was limited to a single dataset in this Aim, we can be more confident in its trajectory identification ability if it is demonstrated in multiple EHR settings.

6.2. Aim 2

In Aim 2, we assessed the ability of CRLI to detect meaningful trajectories in a high-dimensional, multimodal, prospective dataset. In the context of real-world biomedical research data exploration, this represented another commonly utilized longitudinal data source, but one less hampered by the issues we came across in the EHR. Our chosen dataset, the Adolescent Brain Cognitive Development (ABCD) Study, is a longitudinal observational cohort with a prespecified assessment protocol, including a consistent follow-up schedule and a high retention rate (98.9%).[167] This dataset allowed us to explore physical health trajectories (pubertal hormones, anthropometrics) as we did in Aim 1, but also mental health trajectories, as measured by 8 Child Behavior Checklist (CBCL) syndrome scales.[207] We were also able to calculate cluster associations with mental health outcomes as measured by Kiddie Schedule for Affective Disorders and Schizophrenia for School-Aged Children (KSADS) to better characterize cluster differences.

6.2.1. Key findings

Using CRLI, we investigated longitudinal patterns in two domains of development central to adolescent transitions: physical and mental health. In a cohort of 2,923 female participants from the ABCD Study, we identified 3 multivariate trajectories of physical development across pubertal hormones (DHEA, testosterone, estradiol) and anthropometrics (BMI calculated from height and weight). In a cohort of 310 participants who exhibited self-injurious behaviors (SIB) at the year 3 ABCD assessment, we identified multivariate trajectories of psychopathology (8 CBCL syndrome scales) in the years leading up to the SIB. We conducted association testing between cluster membership and outcomes at year 3 and year 4, revealing several significant associations with KSADS symptoms and diagnoses. In the context of the overarching question of how and when IMVTS deep clustering methods are useful in biomedical data exploration, the most important findings were that (1) given longitudinal and static variables, CRLI identified longitudinal

trajectories that had non-uniform associations with static variables, providing a basis for testable clinical hypotheses, and (2) CRLI identified clusters that could not have been identified with a single timepoint or single variable alone. However, our assumption that working with an observational cohort would improve sampling irregularity (due to regularly spaced annual assessment intervals) was violated when we converted the timepoints to age at last birthday (ALB). Working with a dataset where age is less of a biological concern (as it very much is during adolescence) or one with a larger span of timepoints could alleviate this issue.

In Aim 1, we used basic mean lineplots with confidence interval bands to show trajectory clusters. In Aim 2, we augmented our visualizations with boxplots (to more faithfully represent participant spread in each cluster), measurement count tables (to show amount of missingness in each cluster at each timepoint), and average timepoint median barplots (to capture the high-level differences between clusters for each longitudinal variable).

6.2.2. Limitations

In an attempt to extract the longest trajectories possible, our data point selection included year 4, which was incomplete at the time of ABCD Release 5.0. We could have analyzed variables measured at mid-year remote check-ins (4 additional timepoints), but this would have introduced a new hurdle for our chronological age approach (Section 4.4.5.1). The incomplete year 4 also meant that we had fewer participants for which KSADS were recorded, making our cluster association testing less powerful than at year 3. However, at year 3, the KSADS assessment battery itself was limited, so neither timepoint was ideal (Section 4.4.3). Also, as in Aim 1, we faced the challenge of discordant internal CVI and ultimately motivated our cluster number choices based on a holistic view of all four CVI. In the context of the overarching question of how and when IMVTS deep clustering methods are useful in biomedical data exploration, the implications of these limitations are that (1) data availability can impact time series length, which

can in turn affect trajectory discoverability and interpretability, and (2) evaluation of results from unsupervised learning approaches (like clustering) remains difficult because there is definitionally no ground truth that can be compared with.

6.2.3. Future directions

The release of ABCD 6.0 in June 2025 includes complete data for the 4-year follow-up and partial data for years 5 and 6. It also includes improvements to data standardization and documentation, including variable measurement levels (nominal, ordinal, interval, ratio) and an R package for summary score calculation.[268] Of note, gender identity data was omitted from this release, and presumably future releases, due to changing agency priorities.[269] Nevertheless, trajectory subgroup analyses will benefit from more timepoints becoming available and lead to better pattern discovery and more discernable subpopulations. Additionally, some important longitudinal modalities, like imaging, neurocognition, and actigraphy (Fitbit), were underexplored in our work. Unimodal clusterings of each, or multimodal clusterings that bring together diverse data types, will be important to paint a fuller picture of developmental transitions. In the context of exploring high-dimensional biomedical datasets, demonstrating CRLI utility in the multimodal aspect will greatly expand the kinds of clinical questions that can be asked and potentially answered. In the context of analyzing prospective cohorts, specifically ones where we expect to find trajectory divergence at specific timepoints, re-analyzing later releases of ABCD will substantiate the added value that each successive timepoint brings. For example, in Figure 46 we observe that cluster trajectories began to meaningfully diverge only in the last 1 or 2 timepoints (ages 13-14). Re-running the same analysis with the addition of ages 14-15 and 15-16 may yield more separable trajectories.

In Section 4.4.5.1, we discussed our approach to lining up participant time series by chronological age instead of assessment timepoint. For clinical interpretability, we converted participant age,

recorded in months in ABCD, to age last birthday (ALB). Future work could explore increasing this granularity to half years, quarter years, or even keeping age in months. Though this would increase granularity (time series length and sampling regularity), it would also increase missingness, so whether it would lead to more informative or discernible trajectories is unknown. In the context of exploring real-world biomedical datasets, this kind of analysis can formalize the amount of missingness CRLI can tolerate before its ability to detect distinct trajectories is degraded. A comparison between clustering results from converted and unconverted timepoints could inform best practices for working with observational cohorts where age of participant is of consequence to trajectory interpretability.

Several groups have employed model-based clustering methods to capture latent trajectory subgroups in ABCD. This presents an opportunity to feed the same longitudinal variables into CRLI, or another advanced deep learning method, to see if the same subgroups are found across methods. Comparative analyses (model-based vs. algorithm-based clustering) have been performed in other contexts, but not including IMVTS methods and not on ABCD data specifically.[35,60] These analyses can be combined with a wider-ranging exploration of cluster associations with both baseline features and outcomes. This could take the form of a classification model or phenome-wide association study (PheWAS), as some groups have performed on ABCD recently, to see which static features are most important to discern clusters from each other.[241,242] In the context of biomedical research data exploration, comparative analyses using the same dataset(s) can demonstrate which method, or group of methods, is most suitable for a specific dataset or type of data source. Testing associations with high dimensional risk factors and outcomes can provide important context for deriving clinical meaning of clusters, beyond analysis of the longitudinal measures themselves.

6.3. Aim 3

In Aim 3, we assessed the ability of CRLI and VaDER to detect meaningful trajectories in simulated datasets under diverse data constraints. In the context of biomedical research data exploration, this provided a (i) robust synthetic complement to the real-world datasets analyzed in Aims in 1 and 2 and (ii) a comprehensive benchmarking of two state-of-the-art MVTS deep clustering methods. Our work in Aims 1 and 2 crystallized a number of MVTS dataset characteristics (time series length, noise, missingness, number of clusters, number of samples per clusters) that we felt needed further exploration in a synthetic context where we could take a targeted approach to investigating each characteristic's impact on clustering performance. We designed a framework using the mockseries Python package that let us rapidly generate many unique MVTS datasets by randomly sampling from a range of values for each characteristic. We also incorporated the ability to modify time series variable properties like trend, rate of change, and seasonality by designing 5 variable styles (linear sinusoidal, linear transition, two linear transitions, U-shaped curve, exponential increase) inspired by biomedical trends we observed in Aims 1 and 2 and the biomedical time series clustering literature. This purely synthetic approach to method performance evaluation can serve as a complement to the real-world dataset battery (drawing from UCR, UCI, UEA time series archives) that has generally served as the standard evaluation benchmark to date for MVTS clustering.

6.3.1. Key findings

We generated 20 MVTS datasets which varied across all of the aforementioned properties (Table 26) and clustered them with VaDER and CRLI. We reported performance on 4 external CVI (purity, RI, ARI, NMI), of which ARI has been neglected by benchmark evaluations of VaDER and CRLI to date, even though it is recommended as the standard for external validation.[88] We found that CRLI was equal to or outperformed VaDER on 19/20 datasets, RandomScenario_seed-91 being the exception (Table 31). ARI was by far the most conservative

of the 4 CVI and only crossed 0.5 in 4/40 evaluations (20 datasets*2 methods). As ARI was not reported by Ma et al. in their evaluation of CRLI, our results cast doubt on whether it is actually as performant as marketed in the original paper.[33] In some cases, we observed incongruence between learned representation 2D visualizations (PCA, t-SNE, UMAP) and cluster assignments. This motivates an ablation analysis, especially of VaDER, to see if the 2D mapping of the learned representation combined with a basic clustering algorithm (k-means) outperforms the end-to-end method and why that may be. In the context of the overarching question of how and when IMVTS deep clustering methods are useful in biomedical data exploration, the most important findings were that (1) practitioners should be wary of novel methods that do not report performance on adjusted metrics (ARI, AMI), (2) 2D visualizations are an invaluable interpretability tool, especially when there are too many longitudinal variables to understand on an individual basis, and (3) while CRLI generally outperforms VaDER, neither method achieved across-the-board ARI dominance (Section 5.5.1).

6.3.2. Limitations

Due to computational constraints, we generated only 20 datasets through random sampling (Figure 55), though an exhaustive search of all possible combinations of property values from Table 21 would yield well over 100,000 datasets. For the same reason, we did not perform hyperparameter tuning for either method, the drawbacks of which are discussed in Section 6.5.1.1. Lastly, we designed only 7 variable styles (5 original + 2 combinations), though there are conceivably many more that would be valuable for a comprehensive benchmark. We were limited by manual design in this work, but future approaches could incorporate realistic simulations (empirical data-informed) to bolster the number of variable styles available.[246] In the context of the overarching question of how and when IMVTS deep clustering methods are useful in biomedical data exploration, the implications of these limitations are that (1) our simulation approach was non-exhaustive, but reveals the next steps for which properties are most worth

prioritizing in further methods stress testing, and (2) comprehensive testing of any deep learning method is extremely computationally intensive since any hyperparameter could potentially have non-negligible effects on performance.

6.3.3. Future directions

A number of dataset properties not covered in this Aim should be paid attention in future work: imbalanced clusters, temporal autocorrelation, missingness type (MAR, MNAR), outliers, and possibly others we have not considered. Given the outperformance of CRLI by VaDER when trained on a large sample (7000 data points, Table 26), similarly large datasets should be generated with other properties varied to better assess the performance potential of VaDER. One property that we did not quantify or report is a priori cluster separability; a per-variable, per-dataset quantification of this would capture an additional measure of difficulty that will make comparing datasets easier. Lastly, we can use basic machine learning to understand which properties are most impacting performance by regressing one or more external CVI on dataset properties and using feature importance methods, like SHAP, to rank and visualize impact.[265] This will be especially insightful after generating more datasets with a greater spread of properties. In the context of the overarching question of how and when IMVTS deep clustering methods are useful in biomedical data exploration,(1) generation of more datasets with a greater spread of properties, coupled with (2) systematic (quantitative) analysis of which specific properties most impact performance, can inform best practices for methods usage on novel real-world datasets.

6.4. Cross-cutting observations

In this section, we discuss limitations and future directions that emerged during our exploration of how and when one-stage MVTS clustering methods (VaDER, CRLI) are useful in biomedical research data exploration. This work deeply investigated methods developed at the junction of MVTS imputation, representation learning, and clustering (Figure 78). Though we encountered limitations related to computational constraints, cluster validation, and generalizability, our results serve as a jumping off point for several interesting research directions, some of which are already under investigation.

6.4.1. Limitations

6.4.1.1. Hyperparameter tuning

Deep clustering (DC) methods can be distinguished by their architectural differences (feedforward autoencoder vs. convolutional autoencoder) and their type of clustering objective (flat, hierarchical, non-redundant). Inherent to these structural differences are a number of hyperparameters (number of neurons, type of optimizer, learning rate) that should ideally be tuned for each clustering task individually.[270] However, such optimization is often computationally expensive, and more difficult in the unsupervised setting due to lack of a training set, so IMVTS DC practitioners may resort to using the default hyperparameters reported in a given method's publication, especially when performing comparative evaluations between multiple methods.[32,33,100,271] Others may perform hyperparameter optimization for a task and proceed with the set of hyperparameters that yielded the best performance.[30,31] Depending on the complexity of the structure, an optimally tuned DC method can demonstrate performance gains over an untuned version.[28,29] Though we did not perform hyperparameter tuning, future work should incorporate some amount of tuning into the overall workflow, if computationally viable.

6.4.1.2. Cluster validation indices

6.4.1.2.1. Internal

Internal cluster validation indices (ICVI), which are used to establish clustering acceptability in the absence of ground truth labels, have been studied extensively, with no consensus on any one index that outperforms all others in every clustering scenario.[85,86,88,149,272,273] Liu et al. investigated 11 ICVI across 5 aspects (monotonicity, noise, density, subclusters, skewed distributions) and found that S_Dbw index performed well in all 5.[84] However, Arbelaitz et al. concluded that, while Silhouette index obtained the best results in many of the the contexts they tested, no single ICVI of the 30 tested, including S_Dbw index, showed a clear advantage in every context. They also observed that noise and cluster overlap had a large impact on ICVI performance.[86] Van Craenendonck & Blockeel concluded that all 4 of the ICVI they studied exhibited some undesired properties.[149] Noting this lack of agreement across studies, Hämäläinen et al. recommended using multiple ICVI when performing cluster analysis, as we did in our work.[85]

6.4.1.2.2. External

Of the many available external CVI, which are used to assess clustering performance in the presence of ground truth labels, Adjusted Rand Index (ARI) is generally considered to be the standard.[88] Even so, it is (1) not reported in many novel clustering method evaluations and (2) still a global measure that does not provide a sense of how “good” each individual cluster is. Our experiments in Aim 3 demonstrated that, even when the external CVI for a clustering result is low, we can visually identify that certain clusters are more well-separated than others. This can be quantified with per-cluster metrics, like purity, or simple confusion matrices.[274,275] While we (1) reported ARI in addition to other commonly reported external CVI and (2) attempted to qualitatively (visually) assess per-cluster quality in some of our Aim 3 experiments, our results did not include any quantitative per-cluster metrics.

6.4.1.3. Risks of subtyping in real-world data and cohort generalizability concerns

The greater heterogeneity among real-world patients, compared to trial participants subject to specific eligibility criteria, motivates precision medicine analytical approaches, including subgroup discovery. Segal et al. discusses subgroup analysis in the context of heterogeneous treatment effect (HTE) in real-world data (RWD). They distinguish between 4 objectives that researchers may have when investigating HTE, which can be extended more broadly to apply to biomedical data subgrouping in general: (1) confirming that subgroup differences exist, (2) describing the magnitude and nature of those differences, (3) discovering subgroups through exploratory analyses, and (4) predicting individual-level differences. Specific to objective (3), they underscore the need for principled approaches to interpreting differences between newly discovered subgroups when biological plausibility or social rationale for the differences is not strong. They call for the development of a framework for determining whether evidence gathered from subgroup analysis is clinically actionable.[276]

In Aims 1 and 2, our final cohorts were much smaller than the original dataset sample they drew from. This raises questions about generalizability of the clusters identified in the cohort to the population as a whole. Future work should better characterize how the final cohort differed from the dataset sample and how similar either of those is to the actual target population. The same kind of discrepancy that can exist between clinical trial internal and external validity can be observed in our RWD context. Stuart et al. discuss emerging methods for assessing and enhancing trial external validity, like propensity scores, principles of which can be borrowed from to inform RWD analyses as well.[277]

6.4.1.4. Variable-length time series

VaDER and CRLI can handle time series of unequal length by padding missing values onto shorter sequences and imputing them, but they still require a fixed-length grid that all input

sequences must adhere to. However, some MVTS deep clustering methods are able to natively ingest variable-length sequences by leveraging properties of the recurrent neural network (RNN), representative architectures of which are Long Short Term Memory (LSTM) and Gated Recurrent Units (GRU).[27] Trosten et al. introduced Recurrent Deep Divergence-based Clustering (RDDC), which builds on Deep Divergence-based Clustering (DDC) with a two-layer bidirectional GRU, to cluster multivariate sequential data with variable sequence length.[278] Ienco & Interdonato published constrained Deep embedding time SEries Clustering (conDetSEC), a semi-supervised deep embedding time-series clustering framework that is also based on GRUs and can take varying length inputs.[279] Notably, neither of these methods incorporate imputation into their pipelines, so they are not directly comparable to VaDER and CRLI in that regard.

6.4.2. Recent advances and future directions

6.4.2.1. Novel one-stage methods

After the design and execution of this dissertation, two papers by Li et al. introduced novel one-stage deep clustering methods and compared their performance against VaDER and CRLI. The first (2024) introduced an End-to-End Deep Fuzzy Clustering (EEDFC) model and the second (2025) introduced a Multi-View Clustering Method for Incomplete Multivariate Time Series (MVCIMTS).[100,271] EEDFC was compared to eleven other methods: 9 two-stage methods and 2 one-stage methods (VaDER, CRLI). MVCIMTS was compared to 5 incomplete multi-view clustering (IMVC) methods and 2 temporal methods (VaDER, CRLI). Performance was validated via application to a selection of 7-10 benchmark datasets from popular time series dataset repositories (UCR, UCI, UAE).[280–282] The papers also explored method robustness against missingness and noise, reporting evaluation metrics Rand Index (RI), normalized mutual information (NMI), accuracy (ACC), and purity. Future work should consider these methods alongside VaDER and CRLI and include ARI in the evaluation metrics. In the context of the overarching question of how and when IMVTS deep clustering methods are useful in biomedical

data exploration, these methods represent the new state-of-the-art and improve the landscape of available approaches for joint imputation and clustering of longitudinal data.Whether these methods will achieve high ARI performance across diverse synthetic datasets is to be determined.

6.4.2.2. Outcome-guided clustering

In parallel to the purely unsupervised clustering methods discussed in this dissertation, a strain of predictive (outcome-informed) clustering methods has developed to take better advantage of the disease status information that is often available in clinical datasets.[1,283–286] The van der Schaar group has made several contributions to this area, the most recent of which is T-Phenotype (2023), a temporal clustering method which takes advantage of both (i) multivariate longitudinal trends and (ii) static outcomes of interest. They applied their method to synthetic data, observational cohort data (Alzheimer's Disease Neuroimaging Initiative), and ICU data (PhysioNet), demonstrating its prognostic value across multiple dataset types.[283] Wang et al. proposed Outcome-Guided Deep Temporal Clustering (OG-DTC) in 2024, which incorporates an autoencoder and clustering objective like other deep clustering methods, but adds a clinical outcome regression to enhance interpretability and clinical significance. This enables the identification of clusters with homogenous temporal patterns and close outcomes.[285] In the context of the overarching question of how and when IMVTS deep clustering methods are useful in biomedical data exploration, these outcome-informed approaches can used to revisit experiments like those in Aim 2, where specific outcomes of interest are recorded in the dataset (ABCD), with a sharper focus on deriving clinical meaning.

6.4.2.3. Explainability

While deep learning approaches confer many advantages for clustering tasks over traditional approaches, they suffer from a lack of explainability, and, therefore, interpretability. Explainability methods should be able to (1) provide justification for the existence of a given cluster (*do cluster*

members share common patterns?), (2) highlight cluster differences (*what makes a member of cluster A different from a member of B?*), and/or (3) reveal what and where the model architecture is capturing subtleties of cluster differences (*are certain neurons capturing particular patterns?*). In their survey of recent advances in explainable deep clustering for time series data, Schlegel et al. emphasize the accuracy-interpretability tradeoff and motivate several research opportunities in the area. They also reference a proposed taxonomy of explainability techniques: (1) pre-clustering explainability, (2) in-clustering explainability, and (3) post-hoc explainability.[91] Similarly, Huang et al. grouped time series clustering explainability methods into 3 categories: (1) data preprocessing for explainability, (2) model training for explainability, and (3) visualization for explainability. They also provide a list of qualities that ultimately influence a deep clustering model's overall trustworthiness: explainability, interpretability, interactivity, stability, robustness, reproducibility, and confidence.[287] Both papers stress the importance of explainability in the healthcare domain, where clusters must be interpretable for patients and providers to be medically useful. AppendixTable 8, from Schlegel et al., summarizes representative deep time series clustering approaches that incorporate explainability. In the context of detecting meaningful trajectories in biomedical datasets, explainability approaches can help inform what specific meaning should be derived from a clustering result by illuminating which aspects of each cluster the deep learning methods used to actually differentiate them. This quantitative meaning can then be translated into clinical meaning by other domain experts to encourage clinical trustworthiness and actionability.

6.4.2.4. Synthetic clustering benchmarks

Recently, motivated by smart meter time series analyses, Yerbury et al. generated synthetic data to evaluate a range of time series clustering methods (31 distance measures, 8 representations, 11 clustering algorithms) across varying dataset characteristics. They report that performance was most impacted by cluster imbalance, noise, and outliers. However, (1) their investigation did

not include deep learning methods, (2) they did not generate multivariate time series (only univariate), and (3) they did not explore the effect of missingness on model performance.[288] Most similar to our Aim 3 work is the recently published CSTS (Correlation Structures in Time Series), a multivariate synthetic time series benchmark. Their generation framework is customizable, allowing for users to modify dataset characteristics like correlation structure, distribution properties, segment lengths, sparsity, and number of subjects and time series variables generated. However, their generated data lacks temporal dependencies (autocorrelation, trends, seasonality), instead focusing on correlation structures between variables.[289] In the context of assessing MVTS clustering method ability to detect trajectories in synthetic datasets under diverse data constraints, complementary simulation approaches, like the aforementioned, can be integrated with our own to create a comprehensive synthetic benchmark that probes every conceivable angle of method performance. In addition to real-world MVTS repositories, this benchmark could be used by future methods developers to instill confidence in researchers that their methods are robust to wide-ranging data complexities.

6.5. Conclusion

Deep clustering for time series is a rapidly evolving and expanding field. In this dissertation, we focused on applications to longitudinal biomedical data from real-world and synthetic sources which involved an array of data-specific and clinical domain-specific nuances. Across three distinct aims, we explored how and when recently developed one-stage incomplete multivariate time series (IMVTS) deep clustering methods (VaDER, CRLI) are useful in biomedical research data exploration. In Aim 1, we showed that, in the EHR, CRLI can be used to identify post-treatment multivariate response trajectories despite small cohort size (less than 500) and infrequent measurements. In Aim 2, in the observational cohort setting, we demonstrated CRLI's ability to identify multivariate trajectories that had significant differences in proportions of static future outcomes of interest. In particular, we showed that CRLI identified meaningful clusters that

could not have been otherwise identified without multiple timepoints and multiple variables. In Aim 3, we evaluated VaDER and CRLI performance on synthetic datasets using Adjusted Rand Index, which has not been reported to date. We showed that (1) neither method achieved satisfactory ARI performance across 20 diverse datasets and (2) external clustering validation indices (CVI) alone are not sufficient to understand clustering performance. We generated 2D visualizations of the latent representations learned by VaDER and CRLI and analyzed them alongside external CVI to demonstrate the not uncommon disconnect between qualitative (visual) latent space cluster separability and quantitative external CVI performance. Since the initiation of this dissertation, several advances have been made that can augment future extensions of our work. We remain incredibly excited about the potential that this area of research has for better understanding longitudinal patterns in complex biomedical datasets.

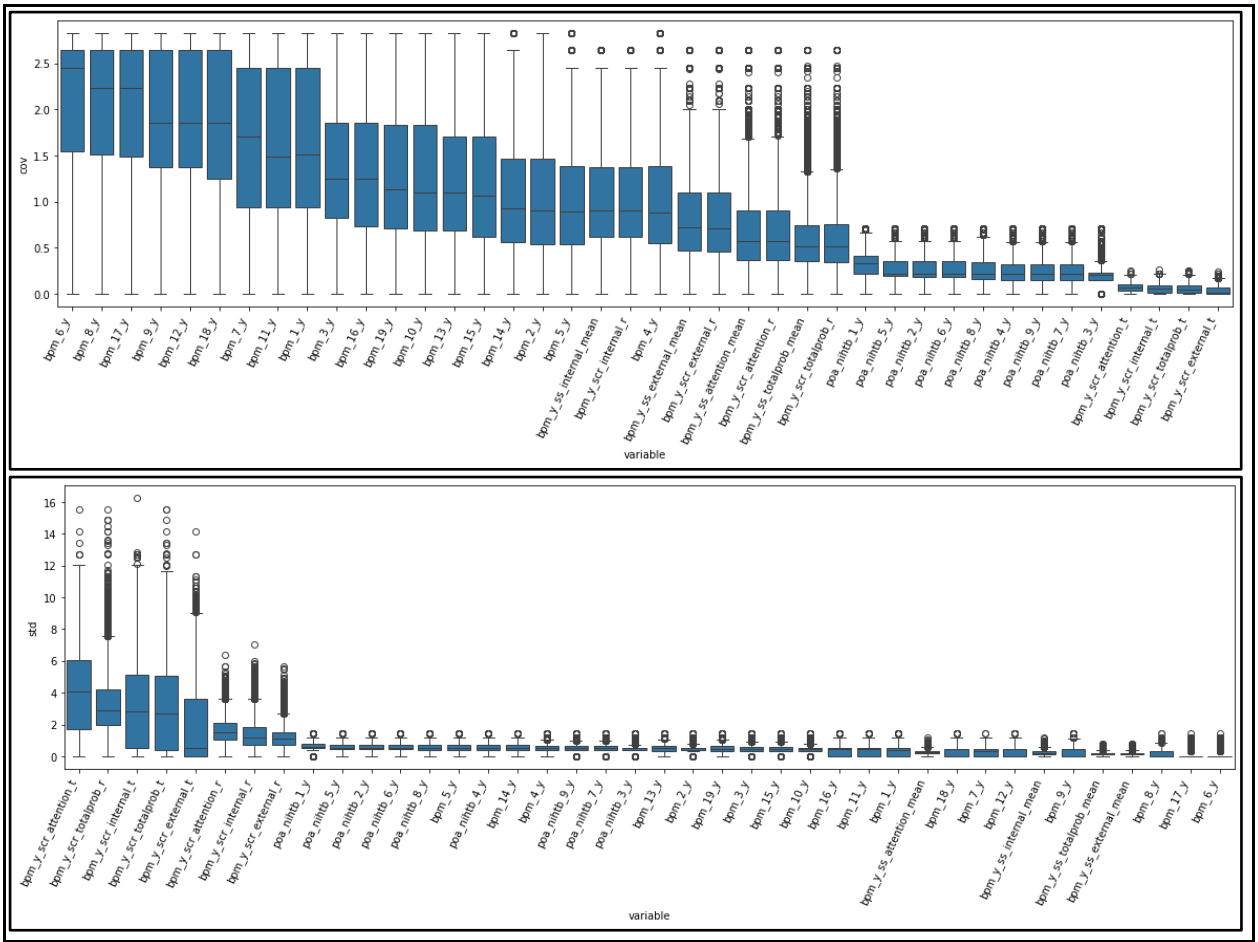
Appendix

Publication	Dataset	Longitudinal variables	Risk factors & Outcomes
Voepel-Lewis 2025[24]	**ABCD**	anxiety, depression, pain, somatic and somnolence (C-PSST)	Early onset substance use
Senger-Carpenter 2025[25]	**ABCD**	Pain, Psychological, and Somatic Symptom (Pain-PSS)	Family adversities, supportive caregiving
Voepel-Lewis 2023[23]	**ABCD**	co-occurring pain, psychological, and sleep disturbance symptom (pain-PSS)	health care utilization
Orri 2018[192]	Québec Longitudinal Study of Child Development	Irritability, depressive/anxious mood	Suicidal ideation & attempts
Gallant 2020[193]	Monitoring Activities of Teenagers to Comprehend Their Habits (MATCH) study	Physical activity, screen time, sleep	
Bendezú 2022[194]	University of Minnesota (UMN) clinics and neighboring areas		Depression, SIB, frontolimbic neural circuity
Zdebik 2019[195]	Quebec Longitudinal Study of Child Development (QLSCD)	Shyness, anxiety, depression	Internalizing problems
Hanson 2019[196]	Birth-to-Twenty Plus Cohort (Bt20+)	Physical activity, sedentary behavior, sleep	SES, mother's schooling, mother's marital status
Murray 2022[197]	Zurich Project on Social Development from Childhood to Adulthood (z-proso) longitudinal cohort study	ADHD, internalising, and externalising symptoms	Bullying, academic achievement, maternal post-natal depression
Zhang 2021[198]	longitudinal study examining psychosocial determinants of growth and development in Anhui Province, China	Life style indicators	Psychopathological outcomes
Carosella 2023[199]	University of Minnesota	stress experience, expression, and physiology (EEP)	Depressive symptoms, SIB

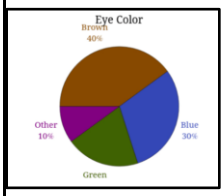
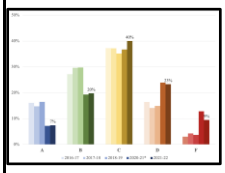
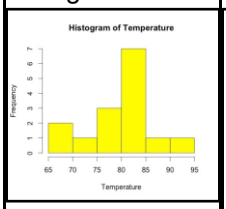
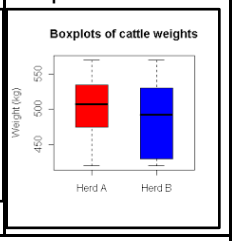
AppendixTable 1. Adolescent trajectory identification in ABCD and non-ABCD datasets

			# of variables after thresholding	
Core assessment domain	**# of tables**	**# of total variables**	**3+ timepoints & 10k+ participants**	**5+ timepoints & 4k+ participants**
ABCD General	5	419	8	8
Culture & Environment	28	515	119	94
Gender Identity & Sexual Health	4	122	1	1
Genetics	2	78	-	-
Imaging	264	56,384	-	-
Linked External Data	71	2,775	-	-
Mental Health	55	9,269	450	291
Neurocognition	16	498	-	-
Novel Technologies	18	2,875	7	3
Physical Health	25	3,364	66	66
Substance Use	41	3,853	17	21
Total	529	80,152	672	488

AppendixTable 2. Summary of longitudinal variables in ABCD Release 5.0



AppendixFigure 1. Variables ranked by median coefficient of variation across all participants and timepoints

Levels of Measurement				
	Qualitative		Quantitative	
	Nominal	**Ordinal**	**Interval**	**Ratio**
Example	Eye color	Letter grades	Temp (F/C)	Weight
Properties				
Category				
Rank				
Equal intervals				
True zero				
Central Tendency				
Mode				
Median				
Arithmetic Mean				
Geometric Mean				
Variability				
Range				
IQR				
SD				
Variance				
CV				
Visualization	Pie 	Bar 	Histogram 	Boxplot 

AppendixTable 3. Levels of measurement[290]

Test name	TB subdomain/construct administration time*	Stimulus and task description	Scores computed For validation study
NIHTB Flanker Inhibitory Control And Attention Test	Executive Attention 6 minutes	Visual display of central arrow pointing left or right, flanked by arrows in the same or opposite direction as the central arrow Task: Indicate direction of central stimulus (leftward or rightward pointing) when flankers are in the same (congruous) or in opposite (incongruous) directions from the central stimulus	Total Trials = 40 Total Score = 10 Score is based on an algorithm derived from both accuracy and reaction time if the former is >80%. If less than 80%, score is accuracy
NIHTB Dimensional Change Card Sort Test	Executive Category Switching 7 minutes	Visual display of two different stimuli side-by-side, each in a different color. Test stimulus matching one of the two display stimuli in either color or shape appears at the bottom of the screen Task: On some trials, shape is the sorting criterion, on others, color. There are 5 pre-switch trials (one category), 5 post-switch trials (after shift to second category) and 30 mixed category trials (shifting)	Total Trials = 40 Total Score = 10 Score is based on an algorithm derived from both accuracy and reaction time if the former is >80%. If less than 80%, score is accuracy
NIHTB List Sorting Working Memory Test	Working Memory: information holding and manipulation 7 minutes	Stimuli from a single category (animals) or two categories (fruits and animals) are presented sequentially visually and aurally in series ranging from 2 to 8 items Task: Orally repeat the sequence of items in order of size. For two-category items, order by size from one specified category first and then from the second.	Total Items correctly sequenced on the one- and two-category trials, of a possible 28
NIHTB Pattern Comparison Processing Speed Test	Processing speed: number of items completed in a finite amount of time 2 minutes	Pairs of stimuli appear side by side Task: Indicate if the stimuli are the "same" or "not the same"	Total number of correct responses within 90 seconds. Maximum score = 130
NIHTB Picture Sequence Memory Test	Episodic Memory for a sequence of pictured events 9 minutes (depending on length of series)	A series of pictures of people performing acts related to a single theme, but not in any intrinsic order, is presented one at a time on the computer screen. After the last item, the pictures are all "collected" in random array in the center of the screen Task: The respondent must place the pictures in the same demonstrated sequence. There are three trials of learning with presentation of the sequence followed by replication of the sequence in each	Total number of correct placements across three learning trials (total possible = 48)
NIHTB Oral Reading Recognition Test	Language: Written word pronunciation 6 minutes	Single printed words are presented in the center of the screen to be read aloud by the respondent. Examiner enters if response is correct or not. Items are presented via CAT method based on the participant's responses	Theta score based on IRT
NIHTB Picture Vocabulary Test	Language: Auditory word-visual picture matching 4 minutes	Four pictures are presented in a two-by-two array on the screen. A single recorded word is presented aurally and the participant must indicate which of the pictures matches the word	Theta score based on IRT

*Administration times are approximate. The norming version has been shortened to remain within the desired 30 minutes originally planned.
IRT = item response theory; NIHTB = NIH Toolbox.

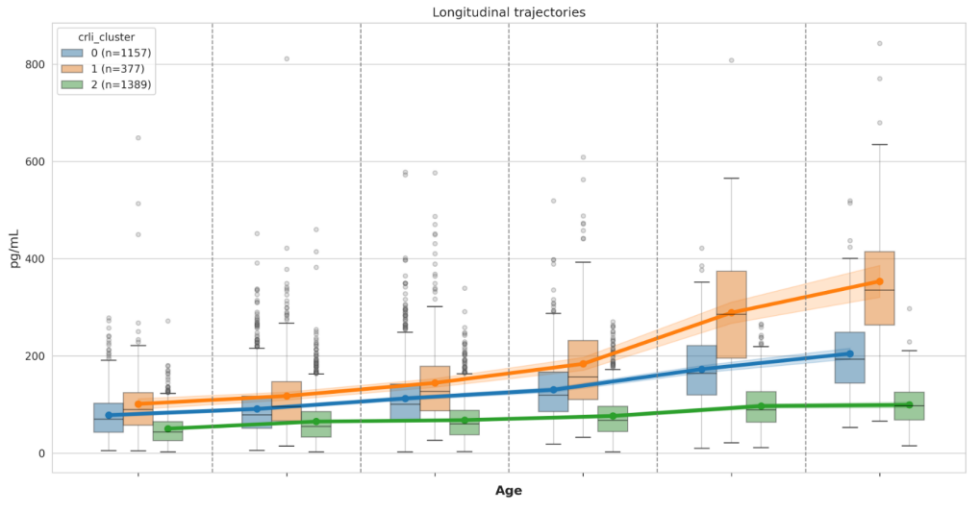
AppendixTable 4. NIH Toolbox Cognition Battery Tests[215]

		Baseline		Year 1		Year 2		Year 3		Year 4		Year 5	
Module #	Module Name	Y	C	Y	C	Y	C	Y	C	Y	C	Y	C
1-3	Mood Disorders	1	1	0	0	1	1	0	0	1	1	0	0
4	Psychosis	0	1	0	1	0	1	0	1	0	1	0	0
5	Panic Disorder	0	1	0	0	0	1	0	0	0	1	0	0
6	Agoraphobia	0	1	0	0	0	1	0	0	0	1	0	0
7	Separation Anxiety	0	1	0	0	0	1	0	0	0	0	0	0
8	Social Anxiety Disorder	1	1	0	0	1	1	0	0	1	1	0	0
9	Specific Phobia	0	1	0	0	0	1	0	0	0	1	0	0
10	Generalized Anxiety Disorder	1	1	0	0	1	1	0	0	1	1	0	0
11	Obsessive Compulsive Disorder	0	1	0	0	0	1	0	0	0	1	0	0
12	Enuresis and Encopresis	0	0	0	0	0	0	0	0	0	0	0	0
13	Eating Disorders	0	1	0	1	1	1	0	1	1	1	0	0
14	Attention Deficit Hyperactivity Disorder	0	1	0	1	0	1	0	1	0	1	0	0
15	Oppositional Defiant Disorder	0	1	0	1	0	1	0	0	0	1	0	0
16	Conduct Disorder	0	1	0	1	1	1	0	1	1	1	0	0
17	Tic Disorders	0	0	0	0	0	0	0	1	0	1	0	0
18	Autism Spectrum Disorders	0	1	0	0	0	1	0	0	0	1	0	0
19	Alcohol Use Disorder	0	1	1*	0	1*	1	1*	0	1*	1	1*	1
20	Drug Use Disorders	0	1	1*	0	1*	1	1*	0	1*	1	1*	1
21	Post-Traumatic Stress Disorder	0	1	0	0	0	1	0	0	0	1	0	0
22	Sleep Problems	1	1	0	0	1	1	0	0	1	1	0	0
23	Suicidality	1	1	1	0	1	1	1	0	1	1	1	0
24	Homicidality	0	1	0	0	0	1	0	0	0	1	0	0
25	Selective Mutism	0	0	0	0	0	0	0	0	0	0	0	0

Note: Y = Youth; C = Caregiver; *Only administered if youth reported substance use.

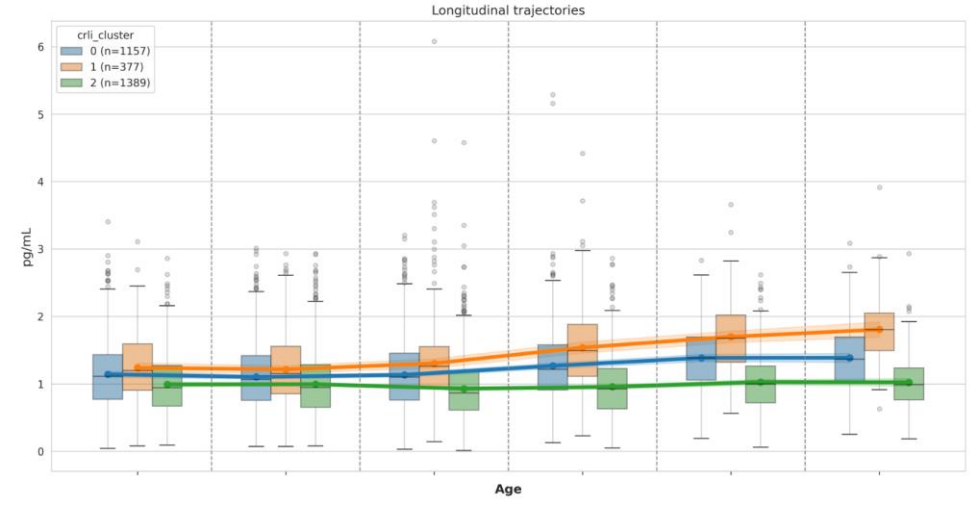
AppendixTable 5. KSADS-COMP modules administered at each wave[182]

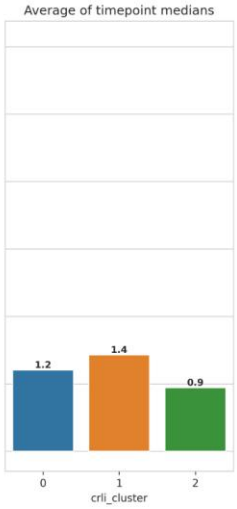
hormone_scr_dhea_mean



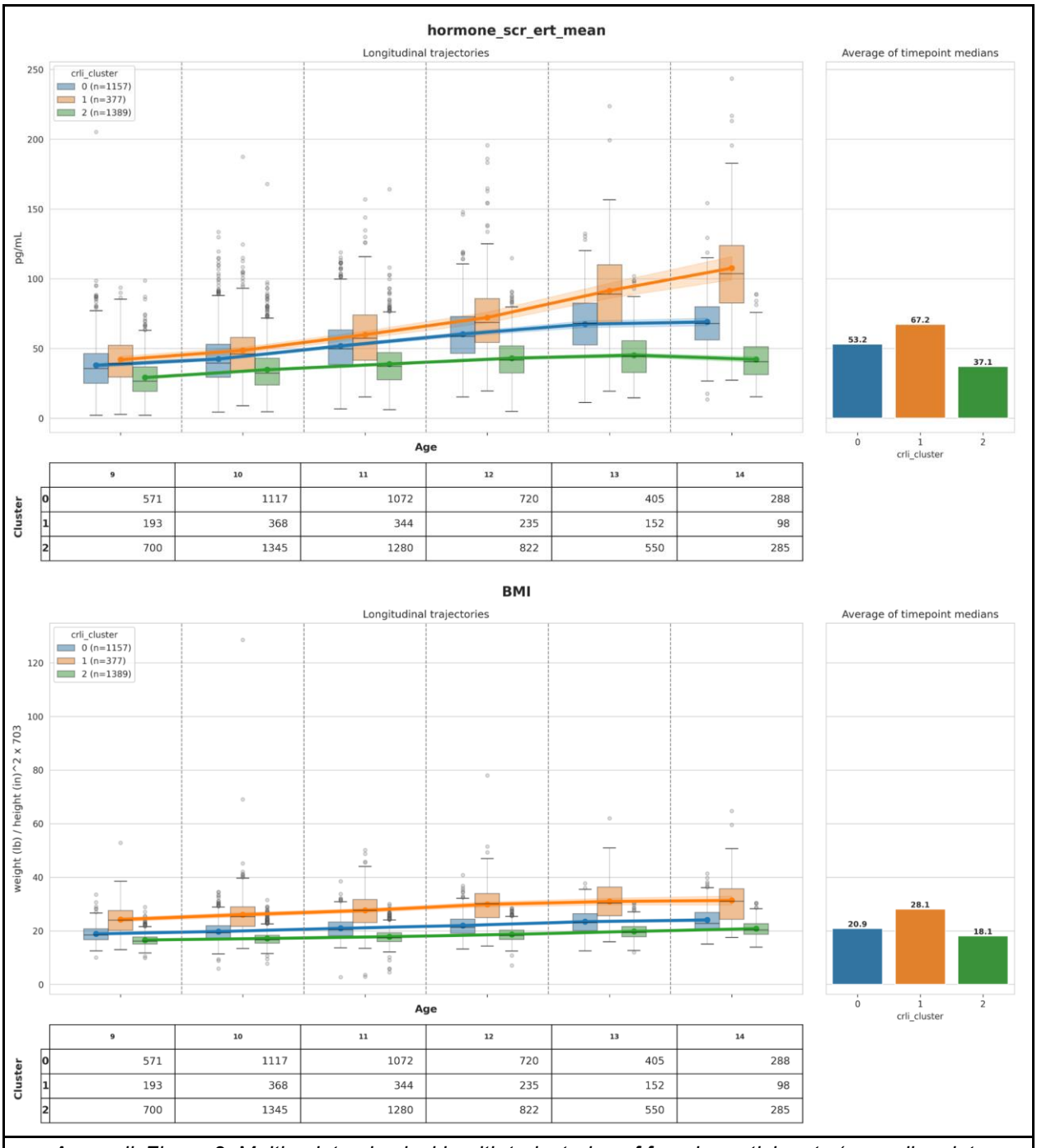
Cluster	9	10	11	12	13	14
0	571	1117	1072	720	405	288
1	193	368	344	235	152	98
2	700	1345	1280	822	550	285

hormone_scr_hse_mean





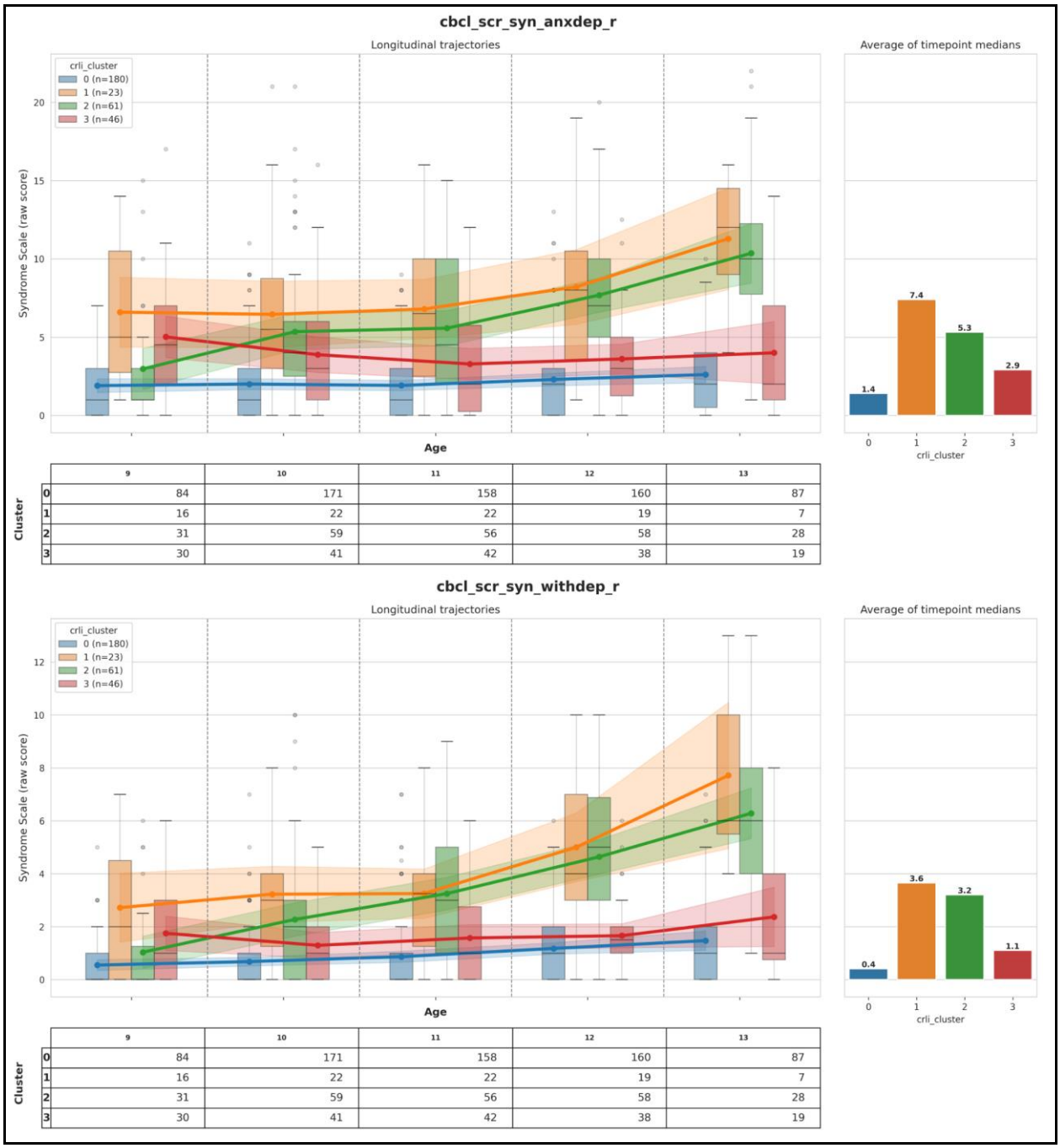
Cluster	9	10	11	12	13	14
0	571	1117	1072	720	405	288
1	193	368	344	235	152	98
2	700	1345	1280	822	550	285



Cluster	9	10	11	12	13	14
0	571	1117	1072	720	405	288
1	193	368	344	235	152	98
2	700	1345	1280	822	550	285

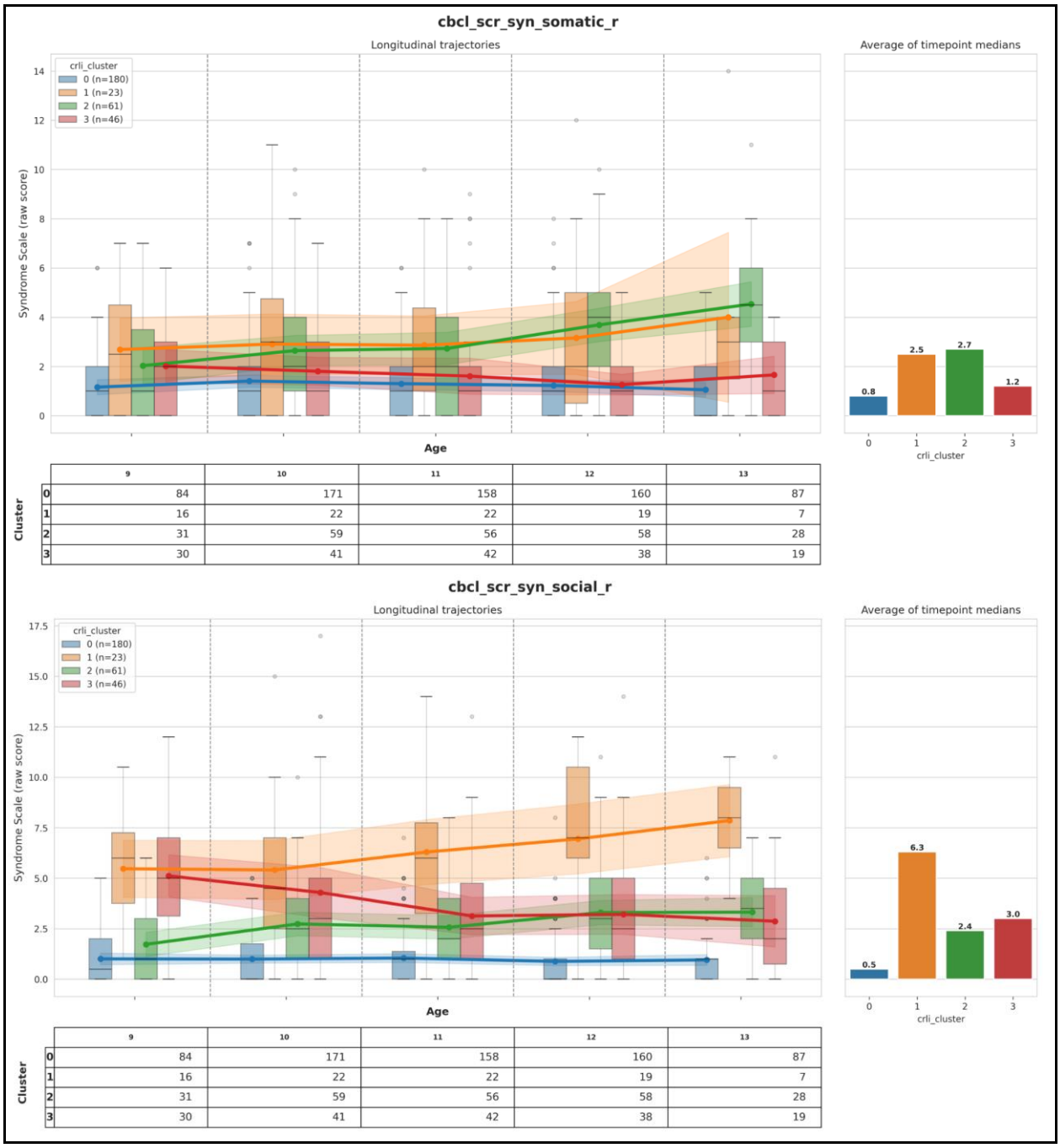
Cluster	9	10	11	12	13	14
0	571	1117	1072	720	405	288
1	193	368	344	235	152	98
2	700	1345	1280	822	550	285

AppendixFigure 2. Multivariate physical health trajectories of female participants (mean lineplots overlaid)



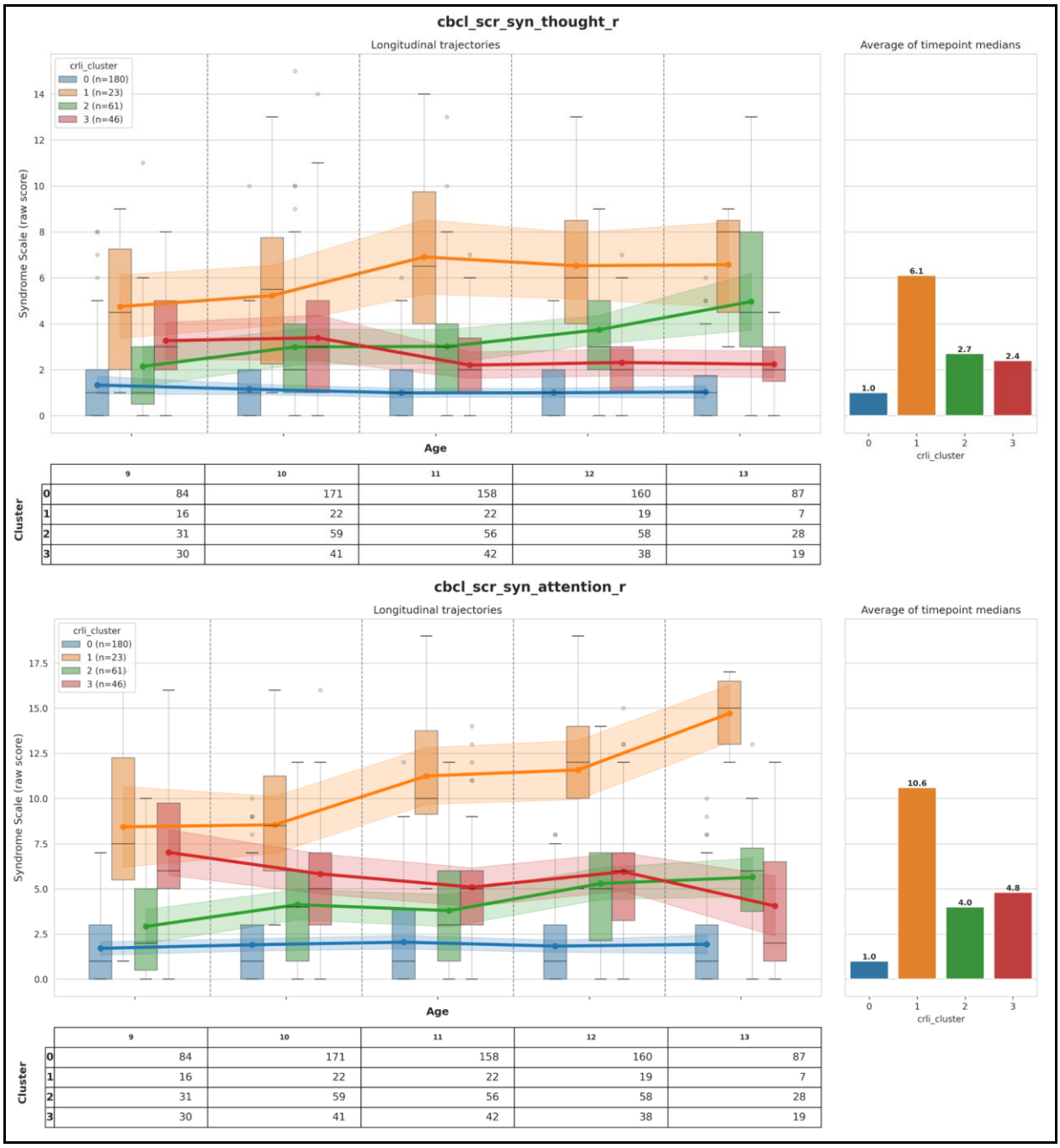
Cluster	9	10	11	12	13
0	84	171	158	160	87
1	16	22	22	19	7
2	31	59	56	58	28
3	30	41	42	38	19

Cluster	9	10	11	12	13
0	84	171	158	160	87
1	16	22	22	19	7
2	31	59	56	58	28
3	30	41	42	38	19



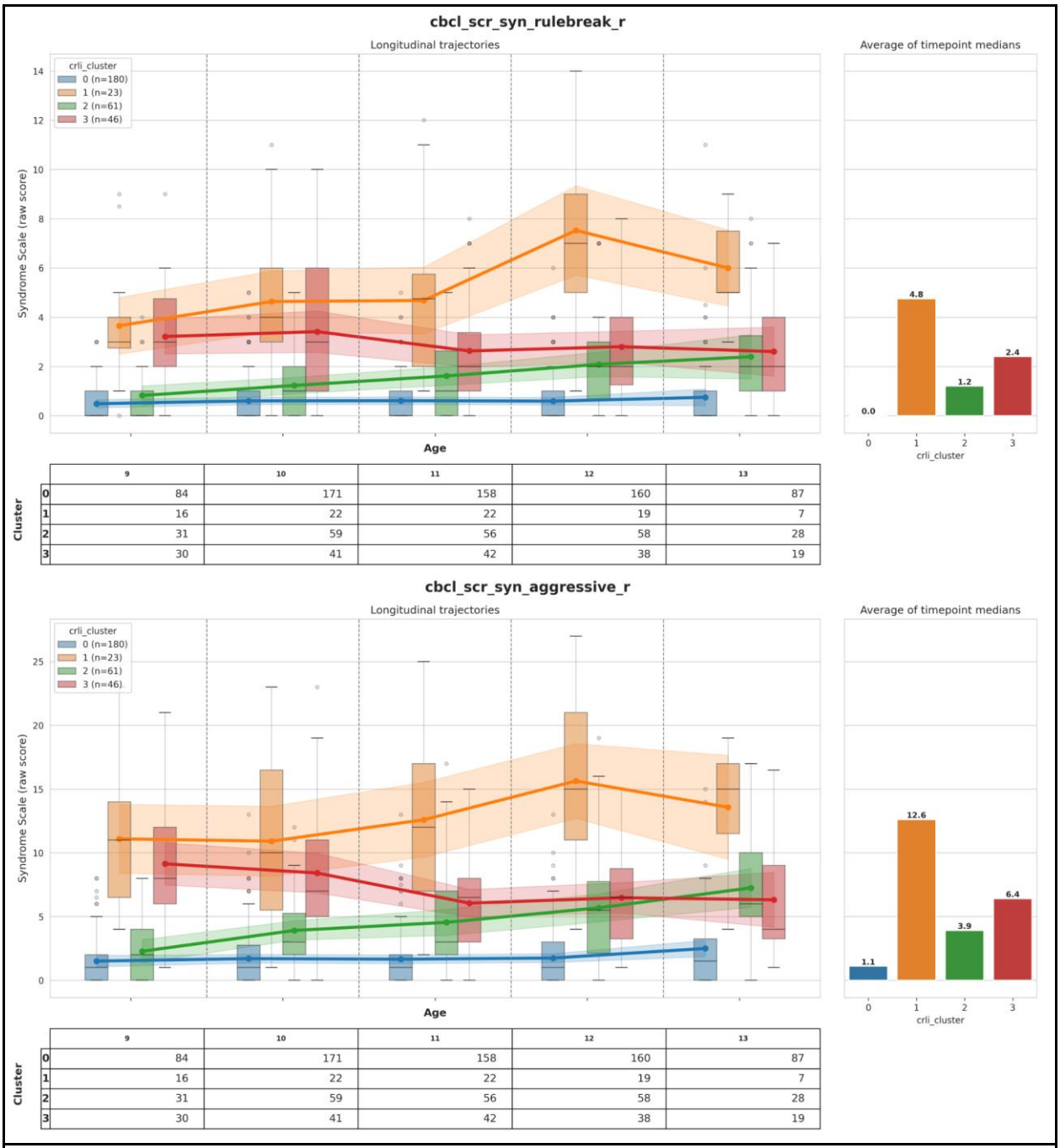
Cluster	9	10	11	12	13
0	84	171	158	160	87
1	16	22	22	19	7
2	31	59	56	58	28
3	30	41	42	38	19

Cluster	9	10	11	12	13
0	84	171	158	160	87
1	16	22	22	19	7
2	31	59	56	58	28
3	30	41	42	38	19

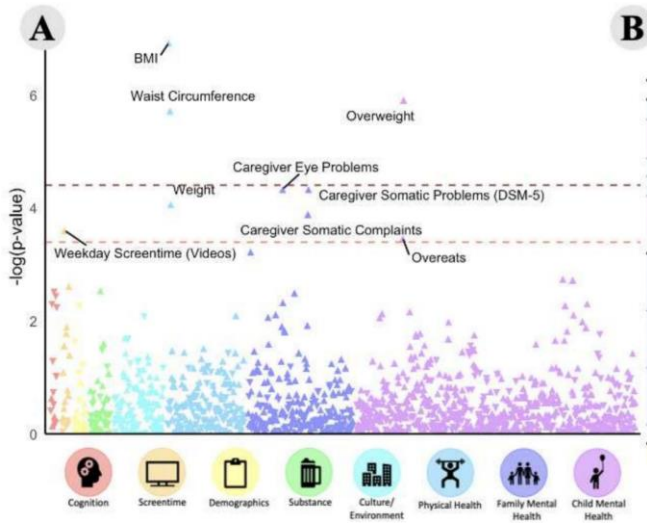


Cluster	9	10	11	12	13
0	84	171	158	160	87
1	16	22	22	19	7
2	31	59	56	58	28
3	30	41	42	38	19

Cluster	9	10	11	12	13
0	84	171	158	160	87
1	16	22	22	19	7
2	31	59	56	58	28
3	30	41	42	38	19



AppendixFigure 3. Multivariate CBCL trajectories of SIB-exhibiting participants (mean lineplots overlaid)



B

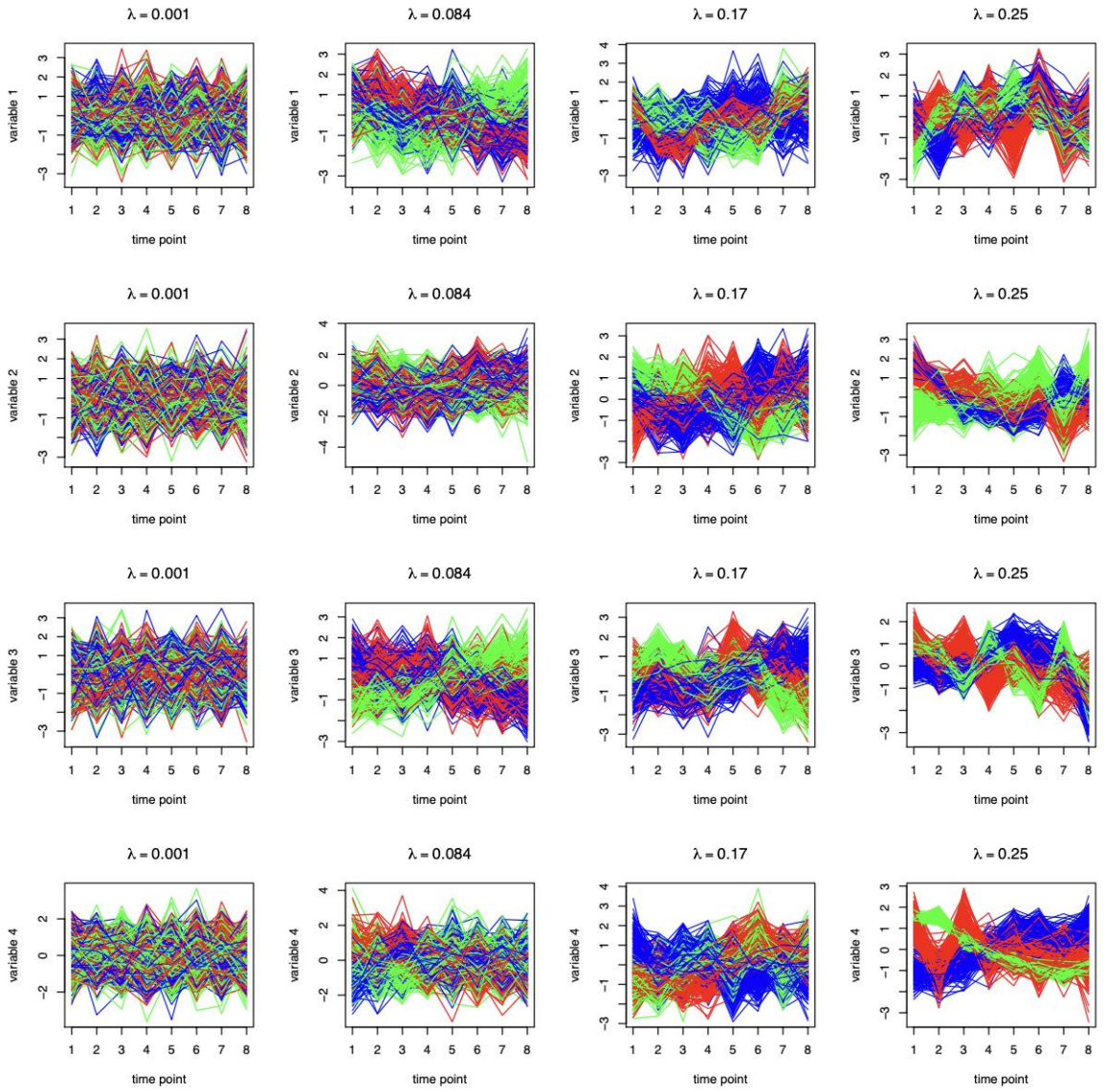
Phenotype description	Survey	Beta	*p* value	FDR corrected *p* value	Bonferroni corrected *p* value
Weight- & Eating-Related Phenotypes:					
Measured BMI	ANT	0.0707	1.20E-07	**0.0002**	**0.0002**
Caregiver report that child is overweight	CBCL	0.0663	1.20E-06	**0.0008**	**0.0016**
Measured waist circumference	ANT	0.0629	1.90E-06	**0.0008**	**0.0024**
Measured weight	ANT	0.0504	8.60E-05	**0.0183**	0.1096
Caregiver report that child overeats	CBCL	0.0487	3.40E-04	**0.0486**	0.4373
Caregiver Somatic Problems:					
Caregiver self-report of eye problems not corrected by glasses	ASR	0.0468	4.60E-05	**0.0119**	0.0587
Caregiver self-report of somatic problems (DSM-5-Oriented Scale)	ASR	0.0451	4.70E-05	**0.0119**	0.0597
Caregiver self-report of somatic problems (ASR Syndrome Scale)	ASR	0.0403	1.30E-04	**0.0231**	0.1619
Other Phenotypes:					
Weekday screentime (videos)	STQ	0.0494	2.50E-04	**0.0394**	0.3155

ANT= Youth Anthropometrics, CBCL = Child Behavior Checklist, ASR = Adult Self-Report, STQ = Youth Screen Time Survey

AppendixFigure 4. Manhattan plot of 1,273 Psychosocial phenotypes[242]

Dataset	Train Cases	Test Cases	Dimensions	Length	Classes
ArticularyWordRecognition	275	300	9	144	25
AtrialFibrillation	15	15	2	640	3
BasicMotions	40	40	6	100	4
CharacterTrajectories	1422	1436	3	182	20
Cricket	108	72	6	1197	12
DuckDuckGeese	60	40	1345	270	5
EigenWorms	128	131	6	17984	5
Epilepsy	137	138	3	206	4
EthanolConcentration	261	263	3	1751	4
ERing	30	30	4	65	6
FaceDetection	5890	3524	144	62	2
FingerMovements	316	100	28	50	2
HandMovementDirection	320	147	10	400	4
Handwriting	150	850	3	152	26
Heartbeat	204	205	61	405	2
JapaneseVowels	270	370	12	29	9
Libras	180	180	2	45	15
LSST	2459	2466	6	36	14
InsectWingbeat	30000	20000	200	78	10
MotorImagery	278	100	64	3000	2
NATOPS	180	180	24	51	6
PenDigits	7494	3498	2	8	10
PEMS-SF	267	173	963	144	7
Phoneme	3315	3353	11	217	39
RacketSports	151	152	6	30	4
SelfRegulationSCP1	268	293	6	896	2
SelfRegulationSCP2	200	180	7	1152	2
SpokenArabicDigits	6599	2199	13	93	10
StandWalkJump	12	15	4	2500	3
UWaveGestureLibrary	120	320	3	315	8

AppendixTable 6. Summary of the 30 datasets in the UEA MVTS classification archive, 2018[247]



AppendixFigure 5. MVTS data simulated using VAR processes (4 variables, 8 time points, 3 clusters) for different levels of similarity parameter λ (from VaDER supplement)[54]

Metric	RMSE		RI		
Method	BRITS	CRLI	Raw_data	BRITS	CRLI
50words	**0.67(0.19)**	1.44(0.29)	0.941(0.01)	**0.952(0.00)**	**0.952(0.00)**
Adiac	**0.93(0.47)**	6.78(2.14)	0.921(0.02)	**0.937(0.00)**	0.934(0.00)
ArrowHead	**1.36(0.50)**	2.01(0.43)	0.556(0.00)	0.503(0.01)	**0.575(0.01)**
Beef	**2.12(1.23)**	2.90(0.99)	**0.651(0.01)**	0.648(0.01)	**0.651(0.00)**
BeetleFly	**1.33(0.39)**	3.03(0.16)	0.479(0.02)	0.516(0.01)	**0.528(0.02)**
BirdChicken	**1.94(1.15)**	3.88(0.85)	**0.474(0.00)**	**0.474(0.00)**	**0.474(0.00)**
Car	**1.24(0.73)**	3.04(1.12)	**0.702(0.05)**	0.678(0.02)	0.698(0.01)
CBF	3.37(0.20)	**3.21(0.08)**	0.704(0.01)	0.699(0.01)	**0.710(0.01)**
Chlorine	**1.15(0.08)**	3.10(0.17)	**0.528(0.00)**	**0.528(0.00)**	**0.528(0.00)**
CinC	**2.19(0.80)**	3.96(0.24)	0.693(0.01)	0.685(0.01)	**0.695(0.00)**
Coffee	**1.11(0.36)**	4.43(1.21)	0.484(0.00)	0.484(0.00)	**0.506(0.01)**
Computers	**5.80(0.37)**	6.40(1.03)	0.502(0.01)	**0.510(0.00)**	0.509(0.00)
Cricket_X	**3.52(0.25)**	4.07(0.60)	0.855(0.01)	0.855(0.00)	**0.859(0.00)**
Cricket_Y	**2.89(0.34)**	3.05(0.22)	0.858(0.01)	0.856(0.00)	**0.861(0.00)**
Cricket_Z	**3.53(0.32)**	4.15(0.89)	0.858(0.01)	0.855(0.00)	**0.863(0.00)**
Diatom	**1.64(0.68)**	2.00(0.38)	0.926(0.00)	0.925(0.00)	**0.927(0.00)**
DPO.AgeGroup	**0.59(0.25)**	1.04(0.04)	**0.764(0.00)**	0.763(0.00)	0.763(0.00)
DPO.Correct	**0.58(0.28)**	0.92(0.05)	**0.505(0.00)**	0.504(0.00)	0.504(0.00)
DP.TW	**0.85(0.50)**	1.07(0.09)	0.774(0.02)	**0.777(0.00)**	0.775(0.00)
Earthquakes	13.02(0.13)	**10.49(0.37)**	0.499(0.03)	0.504(0.01)	**0.511(0.01)**
Average	**2.492**	3.548	0.684	0.683	**0.691**
AVG RANK	**1.1**	1.9	2.2	2.3	**1.4**
***p*-value**	-	-	1.05E-02	3.19E-02	-

AppendixTable 7. RMSE & clustering RI comparison on first 20 synthetic incomplete datasets of UCR archive[33]

Method (Year)	Model	Explainability	Stage	Presented Domain
SOM–VAE (2019) [10]	VAE + 2-D SOM	Latent grid; prototype neurons	In-cluster	ICU vitals; MNIST seq.
T-DPSOM (2020) [21]	VAE + LSTM + prob. SOM	Trajectory map; uncertainty	In-cluster	ICU patient states
DeTSEC (2020) [15]	Attentive GRU autoencoder	Attention on key subsequences	In-cluster	Speech, gesture, ECG
Time2Feat (2022) [6]	Feature extractor + DNN	Interpretable domain features	Pre-cluster	18 IoT / activity sets
CDPS Shapelets (2023) [8]	CNN shapelet learner + k-means	Representative subsequences (shapelets)	Pre-cluster	UCR archive
Explain. EEG Clust. (2023) [9]	Autoencoder + k-means	Cluster-specific spectral patterns	Post-hoc	EEG brain states
CLAMP (2022) [5]	Model-agnostic framework	Post-hoc rules and prototypes	Post-hoc	Cyber-security logs

AppendixTable 8. Representative deep clustering methods for time series with explainability, grouped by the interpretable clustering three-stage taxonomy[91,291]

References

1. Carr, O., Javer, A., Rockenschaub, P., Parsons, O. & Dürichen, R. Longitudinal patient stratification of electronic health records with flexible adjustment for clinical outcomes. **158**, 220–238 (2021).
2. Manzini, E. *et al.* Longitudinal deep learning clustering of Type 2 Diabetes Mellitus trajectories using routinely collected health records. *J. Biomed. Inform.* **135**, 104218 (2022).
3. Xie, Y., Choi, T. & Al-Aly, Z. Mapping the effectiveness and risks of GLP-1 receptor agonists. *Nat. Med.* **31**, 951–962 (2025).
4. Thomsen, R. W., Mailhac, A., Løhde, J. B. & Pottegård, A. Real-world evidence on the utilization, clinical and comparative effectiveness, and adverse effects of newer GLP-1RA-based weight-loss therapies. *Diabetes Obes. Metab.* **27 Suppl 2**, 66–88 (2025).
5. Rodriguez, P. J. *et al.* Semaglutide vs tirzepatide for weight loss in adults with overweight or obesity. *JAMA Intern. Med.* **184**, 1056–1064 (2024).
6. Rodriguez, P. J. *et al.* Discontinuation and reinitiation of dual-labeled GLP-1 receptor agonists among US adults with overweight or obesity. *JAMA Netw. Open* **8**, e2457349 (2025).
7. Zhu, X., Fowler, M. J., Wells, Q. S., Stafford, J. M. & Gannon, M. Predicting responsiveness to GLP-1 pathway drugs using real-world data. *BMC Endocr. Disord.* **24**, 269 (2024).
8. Devineni, D., Akbarpour, M., Gong, Y. & Wong, N. D. Inadequate use of newer treatments and glycemic control by cardiovascular risk and sociodemographic groups in US adults with diabetes in the NIH Precision Medicine Initiative All of Us Research Program. *Cardiovasc. Drugs Ther.* **38**, 347–357 (2024).
9. Mariam-Smith, A. *et al.* Neurobeachin (NBEA) is a novel gene associated with GLP-1 receptor agonist associated weight loss. *Diabetes Obes. Metab.* **27**, 5632–5642 (2025).
10. Mayer, C. S. & Fontelo, P. Semaglutide use in people with obesity and type 2 diabetes from

real-world utilization data: An analysis of the All of US Program. *Diabetes Obes. Metab.* **26**, 4989–4995 (2024).

11. Burke, T. A. *et al.* Nonsuicidal self-injury in preadolescents. *Pediatrics* **152**, (2023).
12. Huber, R. S., Sheth, C., Renshaw, P. F., Yurgelun-Todd, D. A. & McGlade, E. C. Suicide ideation and neurocognition among 9- and 10-year old children in the Adolescent Brain Cognitive Development (ABCD) study. *Arch. Suicide Res.* **26**, 641–655 (2022).
13. Janiri, D. *et al.* Risk and protective factors for childhood suicidality: a US population-based study. *Lancet Psychiatry* **7**, 317–326 (2020).
14. DeVille, D. C. *et al.* Prevalence and family-related factors associated with suicidal ideation, suicide attempts, and self-injury in children aged 9 to 10 years. *JAMA Netw. Open* **3**, e1920956 (2020).
15. Harman, G. *et al.* Prediction of suicidal ideation and attempt in 9 and 10 year-old children using transdiagnostic risk features. *PLoS One* **16**, e0252114 (2021).
16. Ortin-Peralta, A., Sheftall, A. H., Osborn, A. & Miranda, R. Severity and transition of suicidal behaviors in childhood: Sex, racial, and ethnic differences in the adolescent brain cognitive development (ABCD) study. *J. Adolesc. Health* **73**, 724–730 (2023).
17. Wallace, G. T. & Conner, B. T. Longitudinal panel networks of risk and protective factors for early adolescent suicidality in the ABCD sample. *Dev. Psychopathol.* 1–17 (2024).
18. Ho, T. C., Gifuni, A. J. & Gotlib, I. H. Psychobiological risk factors for suicidal thoughts and behaviors in adolescence: a consideration of the role of puberty. *Mol. Psychiatry* **27**, 606–623 (2022).
19. Bendezú, J. J. *et al.* Adolescent cortisol and DHEA responses to stress as prospective predictors of emotional and behavioral difficulties: A person-centered approach. *Psychoneuroendocrinology* **132**, 105365 (2021).
20. Dehestani, N. *et al.* 'Puberty age gap': new method of assessing pubertal timing and its association with mental health problems. *Mol. Psychiatry* **29**, 221–228 (2024).

21. Gracia-Tabuenca, Z., Moreno, M. B., Barrios, F. A. & Alcauter, S. Development of the brain functional connectome follows puberty-dependent nonlinear trajectories. *Neuroimage* **229**, 117769 (2021).
22. Barendse, M. E. A. *et al.* Multimethod assessment of pubertal timing and associations with internalizing psychopathology in early adolescent girls. *J. Psychopathol. Clin. Sci.* **131**, 14–25 (2022).
23. Voepel-Lewis, T. *et al.* Associations of Co-occurring Symptom Trajectories With Sex, Race, Ethnicity, and Health Care Utilization in Children. *JAMA Netw Open* **6**, e2314135 (2023).
24. Voepel-Lewis, T., Stoddard, S. A., Ploutz-Snyder, R. J., Chen, B. & Boyd, C. J. Effect of comorbid psychologic and somatic symptom trajectories on early onset substance use among U.S. youth in the ABCD study. *Addict. Behav.* **160**, 108181 (2025).
25. Senger-Carpenter, T. *et al.* Family adversity and co-occurring pain, psychological, and somatic symptom trajectories from late childhood through early adolescence. *Soc. Sci. Med.* **366**, 117650 (2024).
26. Alqahtani, A., Ali, M., Xie, X. & Jones, M. W. Deep time-series clustering: A review. *Electronics (Basel)* **10**, 3001 (2021).
27. Paparrizos, J., Yang, F. & Li, H. Bridging the gap: A decade review of time-series clustering methods. *arXiv [cs.LG]* (2024).
28. Lafabregue, B., Weber, J., Gançarski, P. & Forestier, G. End-to-end deep representation learning for time series clustering: a comparative study. *Data Min. Knowl. Discov.* **36**, 29–81 (2022).
29. Paparrizos, J. & Bogireddy, S. P. T. R. Time-series clustering: A comprehensive study of data mining, machine learning, and deep learning methods. *Proceedings VLDB Endowment* **18**, 4380–4395 (2025).
30. Birkenbihl, C., de Jong, J., Yalchyk, I. & Fröhlich, H. Deep learning-based patient stratification for prognostic enrichment of clinical dementia trials. *Brain Commun.* **6**, fcae445

(2024).

31. Hähnel, T. *et al.* Progression subtypes in Parkinson's disease identified by a data-driven multi cohort analysis. *NPJ Parkinsons Dis.* **10**, 95 (2024).
32. Wu, A. *et al.* Deep Representation Learning-based dynamic trajectory phenotyping for acute respiratory failure in medical intensive care units. *arXiv [eess.SP]* (2024) doi:10.48550/ARXIV.2405.02563.
33. Ma, Q., Chen, C., Li, S. & Cottrell, G. W. Learning Representations for Incomplete Time Series Clustering. *AAAI* **35**, 8837–8846 (2021).
34. Nagin, D. S., Jones, B. L. & Elmer, J. Recent advances in group-based trajectory modeling for clinical research. *Annu. Rev. Clin. Psychol.* **20**, 285–305 (2024).
35. Lu, Z., Ahmadiankalati, M. & Tan, Z. Joint clustering multiple longitudinal features: A comparison of methods and software packages with practical guidance. *Stat. Med.* **42**, 5513–5540 (2023).
36. Cascarano, A. *et al.* Machine and deep learning for longitudinal biomedical data: a review of methods and applications. *Artif. Intell. Rev.* **56**, 1711–1771 (2023).
37. Spadon, G. *et al.* Pay attention to evolution: Time series forecasting with deep graph-evolution learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **PP**, 1–1 (2021).
38. Trirat, P. *et al.* Universal time-series representation learning: A survey. *arXiv [cs.LG]* (2024) doi:10.48550/ARXIV.2401.03717.
39. Baumgartner, A., Molani, S., Wei, Q. & Hadlock, J. Imputing Missing Observations with Time Sliced Synthetic Minority Oversampling Technique. *arXiv [cs.LG]* (2022).
40. Ma, Q. *et al.* A survey on time-Series Pre-Trained Models. *arXiv [cs.LG]* (2023) doi:10.48550/ARXIV.2305.10716.
41. Middlehurst, M. *et al.* aeon: a Python toolkit for learning from time series. *ArXiv* **abs/2406.14231**, 1–10 (2024).
42. Corporate document. NICE real-world evidennce framework.

https://www.nice.org.uk/corporate/ecd9/resources/nice-realworld-evidence-framework-pdf-1124020816837.

43. Knevel, R. & Liao, K. P. From real-world electronic health record data to real-world results using artificial intelligence. *Ann. Rheum. Dis.* **82**, 306–311 (2023).
44. Liu, F. & Panagiotakos, D. Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Med. Res. Methodol.* **22**, 287 (2022).
45. Real-World Evidence. *U.S. Food and Drug Administration* https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence (2025).
46. Zhao, J., Papapetrou, P., Asker, L. & Boström, H. Learning from heterogeneous temporal data in electronic health records. *J. Biomed. Inform.* **65**, 105–119 (2017).
47. Xie, F. *et al.* Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. *J. Biomed. Inform.* **126**, 103980 (2022).
48. Landi, I. *et al.* Deep representation learning of electronic health records to unlock patient stratification at scale. *NPJ Digit. Med.* **3**, 96 (2020).
49. Den Teuling, N. G. P., Pauws, S. C. & van den Heuvel, E. R. A comparison of methods for clustering longitudinal data with slowly changing trends. *Commun. Stat. Simul. Comput.* **52**, 621–648 (2023).
50. Ployhart, R. E., Bliese, P. D. & Strizver, S. D. Intensive Longitudinal Models. *Annual Review of Organizational Psychology and Organizational Behavior* **12**, 343–367 (2025).
51. Yang, Q. *et al.* Predicting health outcomes with intensive longitudinal data collected by mobile health devices: a functional principal component regression approach. *BMC Med. Res. Methodol.* **24**, 69 (2024).
52. Sun, C., Shen, H., Song, M. & Li, H. A review of deep learning methods for irregularly sampled medical time series data. *arXiv [cs.LG]* (2020).
53. Lu, Z. Clustering longitudinal data: A review of methods and software packages. *Int. Stat.*

Rev. (2024) doi:10.1111/insr.12588.

54. de Jong, J. *et al.* Deep learning for clustering of multivariate clinical patient trajectories with missing values. *Gigascience* **8**, (2019).
55. Mitra, R. *et al.* Learning from data with structured missingness. *arXiv [stat.ML]* (2023) doi:10.48550/ARXIV.2304.01429.
56. Du, W. *et al.* TSI-Bench: Benchmarking Time Series Imputation. *arXiv [cs.LG]* (2024) doi:10.48550/ARXIV.2406.12747.
57. Tan, A. L. M. *et al.* Informative missingness: What can we learn from patterns in missing laboratory data in the electronic health record? *J. Biomed. Inform.* **139**, 104306 (2023).
58. Prada-Ramallal, G., Takkouche, B. & Figueiras, A. Bias in pharmacoepidemiologic studies using secondary health care databases: a scoping review. *BMC Med. Res. Methodol.* **19**, 53 (2019).
59. Gao, C. X. *et al.* An overview of clustering methods with guidelines for application in mental health research. *Psychiatry Res.* **327**, 115265 (2023).
60. Den Teuling, N. G. P. On approaches for clustering longitudinal data: With extensions for modeling therapy adherence of sleep apnea patients. *Research portal Eindhoven University of Technology* https://research.tue.nl/en/publications/on-approaches-for-clustering-longitudinal-data-with-extensions-fo (2023).
61. Nguena Nguefack, H. L. *et al.* Trajectory modelling techniques useful to epidemiological research: A comparative narrative review of approaches. *Clin. Epidemiol.* **12**, 1205–1222 (2020).
62. van der Nest, G., Lima Passos, V., Candel, M. J. J. M. & van Breukelen, G. J. P. An overview of mixture modelling for latent evolutions in longitudinal data: Modelling approaches, fit statistics and software. *Adv. Life Course Res.* **43**, 100323 (2020).
63. Hawes, S. W. *et al.* Longitudinal analysis of the ABCD® study. *Dev. Cogn. Neurosci.* **72**, 101518 (2025).

64. Mikalsen, K. Ø., Bianchi, F. M., Soguero-Ruiz, C. & Jenssen, R. Time series cluster kernel for learning similarities between multivariate time series with missing data. *arXiv [stat.ML]* (2017) doi:10.48550/ARXIV.1704.00794.
65. Warren Liao, T. Clustering of time series data—a survey. *Pattern Recognit.* **38**, 1857–1874 (2005).
66. Aghabozorgi, S., Seyed Shirkhorshidi, A. & Ying Wah, T. Time-series clustering – A decade review. *Inf. Syst.* **53**, 16–38 (2015).
67. Zhou, S. *et al.* A comprehensive survey on Deep Clustering: Taxonomy, challenges, and future directions. *arXiv [cs.LG]* (2022) doi:10.48550/ARXIV.2206.07579.
68. Genolini, C., Alacoque, X., Sentenac, M. & Arnaud, C. kml and kml3d: R Packages to Cluster Longitudinal Data. *J. Stat. Softw.* **65**, 1–34 (2015).
69. Meng, Q. *et al.* Unsupervised representation learning for Time Series: A review. *arXiv [cs.LG]* (2023) doi:10.48550/ARXIV.2308.01578.
70. Wei, X., Zhang, Z., Huang, H. & Zhou, Y. An overview on deep clustering. *Neurocomputing* **590**, 127761 (2024).
71. Understanding LSTM Networks. https://colah.github.io/posts/2015-08-Understanding-LSTMs/.
72. RNN-autoencoder approach for anomaly detection in power plant predictive maintenance systems. *Int. J. Intell. Eng. Syst.* **15**, (2022).
73. Austin. Classify Sentences via a Recurrent Neural Network (LSTM). *Austin G. Walters* https://austingwalters.com/classify-sentences-via-a-recurrent-neural-network-lstm/ (2019).
74. Xie, J., Girshick, R. & Farhadi, A. Unsupervised deep embedding for clustering analysis. *arXiv [cs.LG]* (2015) doi:10.48550/ARXIV.1511.06335.
75. Guo, X., Gao, L., Liu, X. & Yin, J. Improved Deep Embedded Clustering with local structure preservation. in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (International Joint Conferences on Artificial Intelligence Organization,

California, 2017). doi:10.24963/ijcai.2017/243.

76. Jiang, Z., Zheng, Y., Tan, H., Tang, B. & Zhou, H. Variational Deep Embedding: An unsupervised and generative approach to clustering. in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (International Joint Conferences on Artificial Intelligence Organization, California, 2017). doi:10.24963/ijcai.2017/273.
77. Dizaji, K. G., Herandi, A., Deng, C., Cai, W. & Huang, H. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. in *2017 IEEE International Conference on Computer Vision (ICCV)* (IEEE, 2017). doi:10.1109/iccv.2017.612.
78. Mukherjee, S., Asnani, H., Lin, E. & Kannan, S. ClusterGAN: Latent space clustering in generative Adversarial networks. *Proc. Conf. AAAI Artif. Intell.* **33**, 4610–4617 (2019).
79. Ma, Q., Zheng, J., Li, S. & Cottrell, G. Learning representations for time series clustering. *Neural Inf Process Syst* 3776–3786 (2019).
80. Bo, D. *et al.* Structural deep clustering network. in *Proceedings of The Web Conference 2020* (ACM, New York, NY, USA, 2020). doi:10.1145/3366423.3380214.
81. Wang, J. *et al.* Deep learning for multivariate time series imputation: A survey. *arXiv [cs.LG]* (2024) doi:10.48550/ARXIV.2402.04059.
82. Liu, M. *et al.* Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. *arXiv [cs.LG]* (2022) doi:10.48550/ARXIV.2210.08258.
83. Ren, W. *et al.* Moving beyond medical statistics: A systematic review on missing data handling in electronic health records. *Health Data Sci.* **4**, 0176 (2024).
84. Liu, Y., Li, Z., Xiong, H., Gao, X. & Wu, J. Understanding of Internal Clustering Validation Measures. in *2010 IEEE International Conference on Data Mining* 911–916 (2010).
85. Hämäläinen, J., Jauhiainen, S. & Kärkkäinen, T. Comparison of internal clustering validation indices for prototype-based clustering. *Algorithms* **10**, 105 (2017).

86. Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M. & Perona, I. An extensive comparative study of cluster validity indices. *Pattern Recognit.* **46**, 243–256 (2013).
87. Mullin, S. *et al.* Longitudinal K-means approaches to clustering and analyzing EHR opioid use trajectories for clinical subtypes. *J. Biomed. Inform.* **122**, 103889 (2021).
88. Yerbury, L. W., Campello, R. J. G. B., Livingston, G. C., Jr, Goldsworthy, M. & O'Neil, L. On the use of relative Validity Indices for comparing clustering approaches. *ACM Trans. Knowl. Discov. Data* **19**, 1–53 (2025).
89. Javed, A., Lee, B. S. & Rizzo, D. M. A benchmark study on time series clustering. *arXiv [cs.LG]* (2020).
90. Adjustment for chance in clustering performance evaluation. *scikit-learn* https://scikit-learn.org/stable/auto_examples/cluster/plot_adjusted_for_chance_measures.html.
91. Schlegel, U., Tavares, G. M. & Seidl, T. Towards explainable deep clustering for time series data. *arXiv [cs.LG]* (2025) doi:10.48550/arXiv.2507.20840.
92. Zhang, L. *et al.* The medical deconfounder: Assessing treatment effects with electronic health records. *arXiv [stat.ML]* (2019) doi:10.48550/ARXIV.1904.02098.
93. Heumos, L. *et al.* Exploratory electronic health record analysis with ehrapy. *medRxiv* (2023) doi:10.1101/2023.12.11.23299816.
94. Gomes, M. *et al.* Target Trial Emulation for Transparent and Robust Estimation of Treatment Effects for Health Technology Assessment Using Real-World Data: Opportunities and Challenges. *Pharmacoeconomics* **40**, 577–586 (2022).
95. Bica, I., Alaa, A. M., Lambert, C. & van der Schaar, M. From Real-World Patient Data to Individualized Treatment Effects Using Machine Learning: Current and Future Methods to Address Underlying Challenges. *Clin. Pharmacol. Ther.* **109**, 87–100 (2021).
96. Beaulieu-Jones, B. K. *et al.* Examining the Use of Real-World Evidence in the Regulatory Process. *Clin. Pharmacol. Ther.* **107**, 843–852 (2020).
97. Yuanyuan, Z. *et al.* A scoping review of self-supervised representation learning for clinical

decision making using EHR categorical data. *NPJ Digit. Med.* **8**, 362 (2025).

98. Shickel, B., Tighe, P. J., Bihorac, A. & Rashidi, P. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J. Biomed. Health Inform.* **22**, 1589–1604 (2018).
99. Du, W., Yang, Y., Qian, L., Wang, J. & Wen, Q. PyPOTS: A Python toolkit for machine learning on partially-observed time series. *arXiv [cs.LG]* (2025).
100. Li, Y., Du, M., Zhang, W., Jiang, X. & Dong, Y. Feature weighting-based deep fuzzy C-means for clustering incomplete time series. *IEEE Trans. Fuzzy Syst.* **32**, 6835–6847 (2024).
101. Shoop-Worrall, S. J. W. *et al.* Patient-reported wellbeing and clinical disease measures over time captured by multivariate trajectories of disease activity in individuals with juvenile idiopathic arthritis in the UK: a multicentre prospective longitudinal study. *Lancet Rheumatol.* **3**, e111–e121 (2021).
102. Unnikrishnan, A. G. *et al.* Importance of achieving the composite endpoints in diabetes. *Indian J. Endocrinol. Metab.* **17**, 835–843 (2013).
103. Wason, J., McMenamin, M. & Dodd, S. Analysis of responder-based endpoints: improving power through utilising continuous components. *Trials* **21**, 427 (2020).
104. Grayling, M. J. *et al.* Innovative trial approaches in immune-mediated inflammatory diseases: current use and future potential. *BMC Rheumatol* **5**, 21 (2021).
105. Nagin, D. S., Jones, B. L., Passos, V. L. & Tremblay, R. E. Group-based multi-trajectory modeling. *Stat. Methods Med. Res.* **27**, 2015–2023 (2018).
106. Landewé, R. B. M. & van der Heijde, D. Use of multidimensional composite scores in rheumatology: parsimony versus subtlety. *Ann. Rheum. Dis.* **80**, 280–285 (2021).
107. Rudrapatna, V. A. & Butte, A. J. Opportunities and challenges in using real-world data for health care. *J. Clin. Invest.* **130**, 565–574 (2020).
108. Ha, C., Ullman, T. A., Siegel, C. A. & Kornbluth, A. Patients enrolled in randomized

controlled trials do not represent the inflammatory bowel disease patient population. *Clin. Gastroenterol. Hepatol.* **10**, 1002–7; quiz e78 (2012).

109. Ramirez, A. H. *et al.* The All of Us Research Program: Data quality, utility, and diversity. *Patterns (N Y)* **3**, 100570 (2022).

110. All of Us Public Data Browser. https://databrowser.researchallofus.org/.

111. Ross, S. A. A multiplicity of targets: evaluating composite endpoint studies of the GLP-1 receptor agonists in type 2 diabetes. *Curr. Med. Res. Opin.* **31**, 125–135 (2015).

112. Geifman, N. *et al.* Defining trajectories of response in patients with psoriasis treated with biologic therapies. *Br. J. Dermatol.* **185**, 825–835 (2021).

113. Falkenstein, M. J. *et al.* Empirically-derived response trajectories of intensive residential treatment in obsessive-compulsive disorder: A growth mixture modeling approach. *J. Affect. Disord.* **245**, 827–833 (2019).

114. Paul, R. *et al.* Treatment response classes in major depressive disorder identified by model-based clustering and validated by clinical prediction models. *Transl. Psychiatry* **9**, 187 (2019).

115. Parodis, I. *et al.* Trajectories of disease evolution upon treatment initiation in systemic lupus erythematosus: results from four clinical trials of belimumab. *Rheumatology (Oxford)* **64**, 2697–2705 (2025).

116. Anggriani, D. *et al.* Transforming the diabetes mellitus diagnosis and treatment using data technology: Comprehensive analysis of deep learning and machine learning methodologies. *J Sci Ins* **1**, 26–32 (2024).

117. Cardoso, P. *et al.* Phenotype-based targeted treatment of SGLT2 inhibitors and GLP-1 receptor agonists in type 2 diabetes. *Diabetologia* **67**, 822–836 (2024).

118. Salvatore, M. *et al.* Real-world comparative outcomes of GLP-1 RA and semaglutide prescription among individuals with type 2 diabetes. *medRxiv* (2025) doi:10.1101/2025.06.03.25328908.

119. The ‘All of Us’ Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
120. All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).
121. Ter Meer, J. *et al.* A model for rapid innovation for engagement, enrollment, and data and sample collection in a diverse cohort study: Insights from All of Us participant labs. *Mayo Clin. Proc. Digit. Health* **3**, 100227 (2025).
122. Bianchi, D. W. *et al.* The All of Us Research Program is an opportunity to enhance the diversity of US biomedical research. *Nat. Med.* **30**, 330–333 (2024).
123. Klein, D. *et al.* Building a digital health research platform to enable recruitment, enrollment, data collection, and follow-up for a highly diverse longitudinal US cohort of 1 million people in the All of Us Research Program: Design and implementation study. *J. Med. Internet Res.* **27**, e60189 (2025).
124. All of Us Research Program. https://allofus.nih.gov/article/all-us-research-program-protocol.
125. Data Dictionaries. *User Support* https://support.researchallofus.org/hc/en-us/articles/360033200232-Data-Dictionaries.
126. v7 Controlled Tier CDR April 2023 Release Report - Researcher Facing Report. *Google Docs* https://docs.google.com/spreadsheets/d/1VKbgUPCw7k6oRNRB_DJVFLxJoh_alBh_F5E9sA2rbi8/edit?gid=425430088#gid=425430088.
127. OMOP Common Data Model. https://ohdsi.github.io/CommonDataModel/index.html.
128. Mayo, K. R. *et al.* The All of Us data and Research Center: Creating a secure, scalable, and sustainable ecosystem for biomedical research. *Annu. Rev. Biomed. Data Sci.* **6**, 443–464 (2023).
129. Katarzyna Grzeslak, M. A. The Importance of OMOP for Real-World Data. https://www.iqvia.com/blogs/2025/06/the-importance-of-omop-for-real-world-data.
130. Introduction to All of Us Electronic Health Record (EHR) Collection and Data

Transformation Methods. *User Support* https://support.researchallofus.org/hc/en-us/articles/30125602539284-Introduction-to-All-of-Us-Electronic-Health-Record-EHR-Collection-and-Data-Transformation-Methods.

131. WHOCC. ATCDDD - ATC/DDD Index. https://atcddd.fhi.no/atc_ddd_index/?code=A10BJ.
132. All of Us Controlled Tier Dataset v7 CDR Data Dictionary (C2022Q4R13). *Google Docs* https://docs.google.com/spreadsheets/d/1rKV36i0B47WVd9z6dBgO74bCz5bT0rUx1l7hVzFYq0U/edit?gid=1815943286#gid=1815943286.
133. US National Library of Medicine. RxNav. https://mor.nlm.nih.gov/RxNav/search?searchBy=RXCUI&searchTerm=1991306.
134. https://dcricollab.dcri.duke.edu/sites/NIHKR/KR/Using.
135. CKD-EPI Creatinine Equation (2021). *National Kidney Foundation* https://www.kidney.org/ckd-epi-creatinine-equation-2021-0.
136. Changes to eGFR Calculation and What that Means for People Living with Kidney Disease. *National Kidney Foundation* https://www.kidney.org/news-stories/changes-to-egfr-calculation-and-what-means-people-living-kidney-disease.
137. American Diabetes Association Professional Practice Committee. 6. Glycemic targets: *standards of Medical Care in diabetes—2022*. *Diabetes Care* **45**, S83–S96 (2022).
138. Faouzi, J. & Janati, H. Pyts: A python package for time series classification. *J. Mach. Learn. Res.* **21**, 46:1–46:6 (2020).
139. Keogh, E., Chakrabarti, K., Pazzani, M. & Mehrotra, S. Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. *Knowl. Inf. Syst.* **3**, 263–286 (2001).
140. Faouzi, J. *Pyts*. (Github).
141. Data and Statistics Dissemination Policy. *User Support* https://support.researchallofus.org/hc/en-us/articles/22346276580372-Data-and-Statistics-Dissemination-Policy.
142. Piecewise Aggregate Approximation — pyts 0.13.0 documentation.

https://pyts.readthedocs.io/en/latest/auto_examples/approximation/plot_paa.html.

143. Du, W. *PyPOTS: A Python Toolbox for Data Mining on Partially-Observed Time Series*. (Zenodo, 2023). doi:10.5281/ZENODO.6823221.

144. Du, W. *Pypots/clustering/crli/model.py at 1455df3b69e4b6b941ba88ec3b02115c163c998b · WenjieDu/PyPOTS*. (Github).

145. qianlima-lab. *Code/main.py at 890b21a5ac4b0d18b4e1f292ee2e38cd2454d1a7 · Qianlima-lab/CRLI*. (Github).

146. Du, W. *Pypots/clustering/crli/model.py at 064a940297712d1e883ca7c011bfad8a3c5dd522 · WenjieDu/PyPOTS*. (Github).

147. Du, W. *202306_pypots_examples/PyPOTS_Clustering.ipynb at 1209f357c0084698ddfc7fa65340cec8c84f9962 · WenjieDu/BrewPOTS*. (Github).

148. Zheng, Q. *et al.* Deep representation learning from electronic medical records identifies distinct symptom based subtypes and progression patterns for COVID-19 prognosis. *Int. J. Med. Inform.* **191**, 105555 (2024).

149. van Craenendonck, T. & Blockeel, H. Using internal validity measures to compare clustering algorithms. *ICML* 1–8 (2015).

150. Estimated Glomerular Filtration Rate (eGFR). *National Kidney Foundation* https://www.kidney.org/kidney-topics/estimated-glomerular-filtration-rate-egfr.

151. CDC. Adult BMI Categories. *BMI* https://www.cdc.gov/bmi/adult-calculator/bmi-categories.html (2024).

152. Understanding Blood Pressure Readings. *www.heart.org* https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings.

153. Schandelmaier, S. & Guyatt, G. Same old challenges in subgroup analysis-should we do more about methods implementation? *JAMA Netw. Open* **7**, e243339 (2024).

154. Hripcsak, G. *et al.* Characterizing treatment pathways at scale using the OHDSI network. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 7329–7336 (2016).

155. Gratzl, S., Rodriguez, P. J., Cartwright, B. M. G., Baker, C. & Stucky, N. L. Monitoring report: GLP-1 RA prescribing trends - June 2024 data. *medRxiv* (2024) doi:10.1101/2024.01.18.24301500.

156. Xie, Y. *et al.* Comparative effectiveness of SGLT2 inhibitors, GLP-1 receptor agonists, DPP-4 inhibitors, and sulfonylureas on risk of major adverse cardiovascular events: emulation of a randomised target trial using electronic health records. *Lancet Diabetes Endocrinol.* **11**, 644–656 (2023).

157. Khan, S. S., Ndumele, C. E. & Kazi, D. S. Discontinuation of glucagon-like peptide-1 receptor agonists. *JAMA* **333**, 113–114 (2025).

158. Gasoyan, H., Pfoh, E. R., Schulte, R., Le, P. & Rothberg, M. B. Early- and later-stage persistence with antiobesity medications: A retrospective cohort study. *Obesity (Silver Spring)* **32**, 486–493 (2024).

159. Spratt, S. E. *et al.* Assessing electronic health record phenotypes against gold-standard diagnostic criteria for diabetes mellitus. *J. Am. Med. Inform. Assoc.* **24**, e121–e128 (2017).

160. Haneuse, S., Arterburn, D. & Daniels, M. J. Assessing missing data assumptions in EHR-based studies: A complex and underappreciated task. *JAMA Netw. Open* **4**, e210184 (2021).

161. v7 Controlled Tier CDR April 2023 Release Report - Researcher Facing Report. *Google Docs* https://docs.google.com/spreadsheets/d/1VKbgUPCw7k6oRNRB_DJVFLxJoh_alBh_F5E9sA2rbi8/edit?gid=107383282#gid=107383282.

162. American Medical Informatics Association - Unifier: multi-center unit harmonization and quality control pipeline for continuous variable measurements in the All of Us research program. https://amia.secure-platform.com/symposium/gallery/rounds/82001/details/10445.

163. Venkatesh, S. S. *et al.* Characterising the genetic architecture of changes in adiposity during adulthood using electronic health records. *Nat. Commun.* **15**, 5801 (2024).

164. Raman, S. R. *et al.* Analyzing missingness patterns in real-world data using the SMDI toolkit: application to a linked EHR-claims pharmacoepidemiology study. *BMC Med. Res. Methodol.* **24**, 246 (2024).

165. Cronin, R. M. *et al.* Importance of missingness in baseline variables: A case study of the All of Us Research Program. *PLoS One* **18**, e0285848 (2023).

166. Expired NOT-PM-24-003: Request for Information (RFI) on Future Data Linkages within the Center for Linkage and Acquisition of Data for the All of Us Research Program. https://grants.nih.gov/grants/guide/notice-files/NOT-PM-24-003.html?utm_source=chatgpt.com.

167. Feldstein Ewing, S. W. *et al.* Measuring retention within the adolescent brain cognitive development (ABCD)SM study. *Dev. Cogn. Neurosci.* **54**, 101081 (2022).

168. Barch, D. M. *et al.* Demographic, physical and mental health assessments in the adolescent brain and cognitive development study: Rationale and description. *Dev. Cogn. Neurosci.* **32**, 55–66 (2018).

169. Xiang, Q. *et al.* Prediction of the trajectories of depressive symptoms among children in the adolescent brain cognitive development (ABCD) study using machine learning approach. *J. Affect. Disord.* **310**, 162–171 (2022).

170. Duffy, K. A. *et al.* Psychiatric diagnoses and treatment in nine- to ten-year-old participants in the ABCD study. *JAACAP Open* **1**, 36–47 (2023).

171. Brady, K. T., Killeen, T. K., Brewerton, T. & Lucerini, S. Comorbidity of psychiatric disorders and posttraumatic stress disorder. *J. Clin. Psychiatry* **61 Suppl 7**, 22–32 (2000).

172. Paus, T., Keshavan, M. & Giedd, J. N. Why do many psychiatric disorders emerge during adolescence? *Nat. Rev. Neurosci.* **9**, 947–957 (2008).

173. Nagata, J. M., Lee, C. M., Hur, J. O. & Baker, F. C. What we know about screen time and social media in early adolescence: a review of findings from the Adolescent Brain Cognitive Development Study. *Curr. Opin. Pediatr.* **37**, 357–364 (2025).

174. Hagler, D. J., Jr *et al.* Image processing and analysis methods for the Adolescent Brain Cognitive Development Study. *Neuroimage* **202**, 116091 (2019).

175. Overview. https://docs.abcdstudy.org/latest/study/.

176. Hill, E. D. *et al.* Prediction of mental health risk in adolescents. *Nat. Med.* (2025) doi:10.1038/s41591-025-03560-7.

177. Saragosa-Harris, N. M. *et al.* A practical guide for researchers and reviewers using the ABCD Study and other large longitudinal datasets. *Dev. Cogn. Neurosci.* **55**, 101115 (2022).

178. Lopez, D. A. *et al.* Transparency and reproducibility in the Adolescent Brain Cognitive Development (ABCD) study. *Dev. Cogn. Neurosci.* **68**, 101408 (2024).

179. Volkow, N. D. *et al.* The conception of the ABCD study: From substance use to a broad NIH collaboration. *Dev. Cogn. Neurosci.* **32**, 4–7 (2018).

180. Heeringa, S. G. & Berglund, P. A. A guide for population-based analysis of the adolescent brain cognitive development (ABCD) study baseline data. *bioRxiv* (2020) doi:10.1101/2020.02.10.942011.

181. Garavan, H. *et al.* Recruiting the ABCD sample: Design considerations and procedures. *Dev. Cogn. Neurosci.* **32**, 16–22 (2018).

182. Barch, D. M. *et al.* Demographic and mental health assessments in the adolescent brain and cognitive development study: Updates and age-related trajectories. *Dev. Cogn. Neurosci.* **52**, 101031 (2021).

183. Jung, T. & Wickrama, K. A. S. An introduction to latent class growth analysis and growth mixture modeling. *Soc. Personal. Psychol. Compass* **2**, 302–317 (2008).

184. Herting, M. M. *et al.* Correspondence Between Perceived Pubertal Development and Hormone Levels in 9-10 Year-Olds From the Adolescent Brain Cognitive Development Study. *Front. Endocrinol.* **11**, 549928 (2020).

185. Gray, J. C., Schvey, N. A. & Tanofsky-Kraff, M. Demographic, psychological, behavioral,

and cognitive correlates of BMI in youth: Findings from the Adolescent Brain Cognitive Development (ABCD) study. *Psychol. Med.* **50**, 1539–1547 (2020).

186. Adise, S. *et al.* Variation in executive function relates to BMI increases in youth who were initially of a healthy weight in the ABCD Study. *Obesity (Silver Spring)* **31**, 2809–2821 (2023).

187. Elam, K. K., Su, J., Kutzner, J. & Trevino, A. Individual trajectories of depressive symptoms within racially-ethnically diverse youth: Associations with polygenic risk for depression and substance use intent and perceived harm. *Behav. Genet.* **54**, 86–100 (2024).

188. Jia, L. *et al.* Children's early signs and developmental trajectories of psychotic-like experiences. *Brain Res.* **1832**, 148853 (2024).

189. Xiao, Y., Meng, Y., Brown, T. T., Keyes, K. M. & Mann, J. J. Addictive screen use trajectories and suicidal behaviors, suicidal ideation, and mental health in US youths. *JAMA* (2025) doi:10.1001/jama.2025.7829.

190. Karcher, N. R. *et al.* Longitudinal trajectories of cognition and neural metrics as predictors of persistent distressing psychotic-like experiences across middle childhood and early adolescence. *bioRxivorg* (2025) doi:10.1101/2025.01.20.633817.

191. Grimes, P. Z. *et al.* Genetic architectures of adolescent depression trajectories in 2 longitudinal population cohorts. *JAMA Psychiatry* **81**, 807–816 (2024).

192. Orri, M. *et al.* Association of childhood irritability and depressive/anxious mood profiles with adolescent suicidal ideation and attempts. *JAMA Psychiatry* **75**, 465–473 (2018).

193. Gallant, F., Thibault, V., Hebert, J., Gunnell, K. E. & Bélanger, M. One size does not fit all: identifying clusters of physical activity, screen time, and sleep behaviour co-development from childhood to adolescence. *Int. J. Behav. Nutr. Phys. Act.* **17**, 58 (2020).

194. Bendezú, J. J., Thai, M., Wiglesworth, A., Cullen, K. R. & Klimes-Dougan, B. Adolescent stress experience-expression-physiology correspondence: Links to depression, self-injurious thoughts and behaviors, and frontolimbic neural circuity. *J. Affect. Disord.* **300**,

269–279 (2022).

195. Zdebik, M. A. *et al.* Childhood multi-trajectories of shyness, anxiety and depression: Associations with adolescent internalizing problems. *J. Appl. Dev. Psychol.* **64**, 101050 (2019).

196. Hanson, S. K. *et al.* Longitudinal patterns of physical activity, sedentary behavior and sleep in urban South African adolescents, Birth-To-Twenty Plus cohort. *BMC Pediatr.* **19**, 241 (2019).

197. Murray, A. L., Eisner, M., Nagin, D. & Ribeaud, D. A multi-trajectory analysis of commonly co-occurring mental health issues across childhood and adolescence. *Eur. Child Adolesc. Psychiatry* **31**, 145–159 (2022).

198. Zhang, A. *et al.* Joint trajectories of life style indicators and their links to psychopathological outcomes in the adolescence. *BMC Psychiatry* **21**, 407 (2021).

199. Carosella, K. A. *et al.* Patterns of experience, expression, and physiology of stress relate to depressive symptoms and self-injurious thoughts and behaviors in adolescents: a person-centered approach. *Psychol. Med.* **53**, 7902–7912 (2023).

200. ABCD Study. *ABCD Study* https://abcdstudy.org/ (2019).

201. Karcher, N. R. & Barch, D. M. The ABCD study: understanding the development of risk for mental and physical health outcomes. *Neuropsychopharmacology* **46**, 131–142 (2021).

202. National Institute on Drug Abuse. ABCD Curated Data Releases. *National Institute on Drug Abuse* https://nida.nih.gov/research-topics/adolescent-brain/longitudinal-study-adolescent-brain-cognitive-development-abcd-study/abcd-curated-data-releases (2023).

203. NDA. https://nda.nih.gov/abcd/abcd-annual-releases.

204. https://web.archive.org/web/20250201125254/https://wiki.abcdstudy.org/release-notes/start-page.html.

205. Clark, D. A. *et al.* The general factor of psychopathology in the Adolescent Brain Cognitive Development (ABCD) study: A comparison of alternative modeling approaches. *Clin.*

Psychol. Sci. **9**, 169–182 (2021).

206. https://aseba.org/wp-content/uploads/cbclprofile.pdf.

207. Achenbach, T. M., Ivanova, M. Y. & Rescorla, L. A. Empirically based assessment and taxonomy of psychopathology for ages 1½-90+ years: Developmental, multi-informant, and multicultural findings. *Compr. Psychiatry* **79**, 4–18 (2017).

208. Gershon, R. C. *et al.* NIH toolbox for assessment of neurological and behavioral function. *Neurology* **80**, S2–6 (2013).

209. Salsman, J. M. *et al.* Emotion assessment using the NIH Toolbox. *Neurology* **80**, S76–86 (2013).

210. https://github.com/orgs/now-i-know-my-abcd/discussions/70#discussioncomment-9099358.

211. Lacalle Sisteré, M., Domènech Massons, J. M., Granero Pérez, R. & Ezpeleta Ascaso, L. Validity of the DSM-Oriented scales of the Child Behavior Checklist and Youth Self-report. *Psicothema* **26**, 364–371 (2014).

212. Achenbach, T. M. & Rescorla, L. A. Manual for the ASEBA School-Age Forms & Profiles. *Research Center for Children* 16–17 (2001).

213. https://web.archive.org/web/20241211190444/https://wiki.abcdstudy.org/release-notes/non-imaging/physical-health.html#saliva-analysishormones.

214. NIMH Data Archive - Data - Study. https://nda.nih.gov/study.html?id=2147.

215. Weintraub, S. *et al.* The cognition battery of the NIH toolbox for assessment of neurological and behavioral function: validation in an adult sample. *J. Int. Neuropsychol. Soc.* **20**, 567–578 (2014).

216. Weintraub, S. *et al.* Cognition assessment using the NIH Toolbox. *Neurology* **80**, S54–64 (2013).

217. Mental Health. https://wiki.abcdstudy.org/release-notes/non-imaging/mental-health.html#kiddie-schedule-for-affective-disorders-and-schizophrenia-for-school-aged-children-ksads-comp (2025).

218. Kaufman, J. About KSADS-COMP tool and features. https://www.ksadslogin.net/ksads-comp/aboutKC.aspx.
219. https://github.com/orgs/now-i-know-my-abcd/discussions/63.
220. https://abcdstudy.org/wp-content/uploads/2020/11/flyer_protocol-3yrFlup_eg.pdf.
221. https://abcdstudy.org/wp-content/uploads/2022/05/flyer_protocol-4yrFlup_eg-final.pdf.
222. Shore, L., Toumbourou, J. W., Lewis, A. J. & Kremer, P. Review: Longitudinal trajectories of child and adolescent depressive symptoms and their predictors - a systematic review and meta-analysis. *Child Adolesc. Ment. Health* **23**, 107–120 (2018).
223. Dube, S. L., Johns, M. M., Robin, L., Hoffman, E. & Potter, A. S. Comparison of methods to assess adolescent gender identity in the ABCD study. *JAMA Pediatr.* **178**, 86–88 (2024).
224. https://web.archive.org/web/20250215090125/https://wiki.abcdstudy.org/release-notes/non-imaging/mental-health.html.
225. Pollard, T. J., Johnson, A. E. W., Raffa, J. D. & Mark, R. G. tableone: An open source Python package for producing summary statistics for research papers. *JAMIA Open* **1**, 26–31 (2018).
226. statsmodels.stats.multitest - statsmodels 0.15.0 (+841). https://www.statsmodels.org/dev/_modules/statsmodels/stats/multitest.html#multipletests.
227. qianlima-lab. *Code/main.py at Main · Qianlima-lab/CRLI*. (Github).
228. Scientific Python Forum. chi2_contingency — SciPy v1.16.0 Manual. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html.
229. statsmodels.stats.multitest.multipletests - statsmodels 0.15.0 (+661). https://www.statsmodels.org/dev/generated/statsmodels.stats.multitest.multipletests.html.
230. Best Practice — TableOne 0.9.0 documentation. https://tableone.readthedocs.io/en/latest/bestpractice.html#multiple-testing.
231. People at PyPOTS. https://pypots.com/about/ (1AD).
232. Du, W. *PyPOTS*. (Github).

233. https://abcdstudy.org/wp-content/uploads/2023/08/Data_release_schedule.pdf.

234. Compton, W. M., Dowling, G. J. & Garavan, H. Ensuring the best use of data: The adolescent brain cognitive development study. *JAMA Pediatr.* **173**, 809–810 (2019).

235. CDC. Child and Teen BMI Calculator. *BMI* https://www.cdc.gov/bmi/child-teen-calculator/index.html (2025).

236. NIMH Data Archive - Data - Study. https://nda.nih.gov/study.html?id=2313.

237. Paulich, K. N., Ross, J. M., Lessem, J. M. & Hewitt, J. K. Screen time and early adolescent mental health, academic, and social outcomes in 9- and 10- year old children: Utilizing the Adolescent Brain Cognitive Development [SM] (ABCD) Study. *PLoS One* **16**, e0256591 (2021).

238. Newson, J. J., Bala, J., Giedd, J. N., Maxwell, B. & Thiagarajan, T. C. Leveraging big data for causal understanding in mental health: a research framework. *Front. Psychiatry* **15**, 1337740 (2024).

239. Aguinaldo, L. D., Coronado, C., Gomes, D. A., Courtney, K. E. & Jacobus, J. Application of the RDoC framework to predict alcohol use and suicidal thoughts and behaviors among early adolescents in the adolescent brain and cognitive development (ABCD) Study. *Brain Sci.* **12**, 935 (2022).

240. Ralevski, A. *et al.* Using large language models to abstract complex social determinants of health from original and deidentified medical notes: Development and validation study. *J. Med. Internet Res.* **26**, e63445 (2024).

241. Gorelik, A. J. *et al.* A phenome-wide association study (PheWAS) of Late Onset Alzheimer Disease genetic risk in children of European ancestry at middle childhood: Results from the ABCD study. *Behav. Genet.* **53**, 249–264 (2023).

242. Norton, S. A. *et al.* A Phenome-Wide association study (PheWAS) of genetic risk for C-reactive protein in children of European Ancestry: Results from the ABCD study. *Brain Behav. Immun.* **128**, 487–496 (2025).

243. Paul, S. E. *et al.* A phenome-wide association study of cross-disorder genetic liability in youth genetically similar to individuals from European reference populations. *Nat. Ment. Health* **2**, 1327–1341 (2024).
244. Li, M. *et al.* Causal relationships between screen use, reading, and brain development in early adolescents. *Adv. Sci. (Weinh.)* **11**, e2307540 (2024).
245. Boulesteix, A.-L. *et al.* Introduction to statistical simulations in health research. *BMJ Open* **10**, e039921 (2020).
246. Van Mechelen, I. *et al.* A white paper on good research practices in benchmarking: The case of cluster analysis. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **13**, (2023).
247. Bagnall, A. *et al.* The UEA multivariate time series classification archive, 2018. *arXiv [cs.LG]* (2018) doi:10.48550/ARXIV.1811.00075.
248. Time Series Classification Website. https://www.timeseriesclassification.com/.
249. Dau, H. A. *et al.* The UCR time series archive. *IEEE/CAA J. Autom. Sin.* **6**, 1293–1305 (2019).
250. UCI Machine Learning Repository. https://archive.ics.uci.edu.
251. Isasa, I. *et al.* Comparative assessment of synthetic time series generation approaches in healthcare: leveraging patient metadata for accurate data synthesis. *BMC Med. Inform. Decis. Mak.* **24**, 27 (2024).
252. Bock, C., Moor, M., Jutzeler, C. R. & Borgwardt, K. Machine learning for biomedical time series classification: From shapelets to deep learning. *Methods Mol. Biol.* **2190**, 33–71 (2021).
253. Wang, W. K. *et al.* A systematic review of time series classification techniques used in biomedical applications. *Sensors (Basel)* **22**, 8016 (2022).
254. Peyroteo, M., Ferreira, I. A., Elvas, L. B., Ferreira, J. C. & Lapão, L. V. Remote Monitoring Systems for Patients With Chronic Diseases in Primary Health Care: Systematic Review. *JMIR mHealth and uHealth* **9**, e28285 (2021).

255. Poncette, A.-S. *et al.* Clinical requirements of future patient monitoring in the intensive care unit: Qualitative study. *JMIR Med. Inform.* **7**, e13064 (2019).

256. Lasko, T. A., Denny, J. C. & Levy, M. A. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PLoS One* **8**, e66341 (2013).

257. Proios, D., Bornet, A., Yazdani, A., Rodrigues, J. F. & Teodoro, D. ICU-TSB: A benchmark for temporal patient representation learning for unsupervised stratification into patient cohorts. in *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)* 65–70 (IEEE, 2025).

258. Souli, Y., Trudel, X., Diop, A., Brisson, C. & Talbot, D. Longitudinal plasmode algorithms to evaluate statistical methods in realistic scenarios: an illustration applied to occupational epidemiology. *BMC Med. Res. Methodol.* **23**, 242 (2023).

259. de Catheu, C. *Mockseries: Easy and Intuitive Generation of Synthetic Timeseries for Python*. (Github).

260. Du, W. *Pypots/clustering/vader/model.py at 064a940297712d1e883ca7c011bfad8a3c5dd522 · WenjieDu/PyPOTS*. (Github).

261. Arinik, N., Labatut, V. & Figueiredo, R. Characterizing and comparing external measures for the assessment of cluster analysis and community detection. *IEEE Access* **9**, 20255–20276 (2021).

262. Yang, Y. *et al.* Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. *Cell Rep.* **36**, 109442 (2021).

263. McConville, R., Santos-Rodriguez, R., Piechocki, R. J. & Craddock, I. N2D: (not too) deep clustering via clustering the local manifold of an autoencoded embedding. *arXiv [cs.LG]* (2019) doi:10.48550/ARXIV.1908.05968.

264. Bayer, F., Plecko, D., Beerenwinkel, N. & Kuipers, J. Fair Clustering: A Causal Perspective. *arXiv [stat.ML]* (2023).

265. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. *arXiv [cs.AI]* (2017).
266. Parameshwaran, D., Subramaniyam, N. P. & Thiagarajan, T. C. Waveform complexity: A new metric for EEG analysis. *J. Neurosci. Methods* **325**, 108313 (2019).
267. All of Us Research Program. https://allofus.nih.gov/article/announcement-all-of-us-adds-data-from-50-more-participants-in-largest-data-expansion-to-date.
268. 6.0 data release. https://docs.abcdstudy.org/latest/documentation/release_notes/6_0.html.
269. McMurray, C. ABCD Study omits gender-identity data from latest release. *The Transmitter: Neuroscience News and Perspectives* https://www.thetransmitter.org/gender/abcd-study-omits-gender-identity-data-from-latest-release/ (2025).
270. Leiber, C., Miklautz, L., Plant, C. & Böhm, C. Benchmarking deep clustering algorithms with ClustPy. in *2023 IEEE International Conference on Data Mining Workshops (ICDMW)* (IEEE, 2023). doi:10.1109/icdmw60847.2023.00087.
271. Li, Y., Du, M., Jiang, X. & Zhang, N. Contrastive learning-based multi-view clustering for incomplete multivariate time series. *Inf. Fusion* **117**, 102812 (2025).
272. Hassan, B. A. *et al.* From A-to-Z review of clustering validation indices. *Neurocomputing* **601**, 128198 (2024).
273. Liu, Y. *et al.* Understanding and enhancement of internal clustering validation measures. *IEEE Trans. Cybern.* **43**, 982–994 (2013).
274. Chavooshi, M. & Mamonov, A. V. Autoencoded UMAP-Enhanced Clustering for unsupervised learning. *arXiv [cs.LG]* (2025).
275. Lee, A. S. *et al.* A cell type-aware framework for nominating non-coding variants in Mendelian regulatory disorders. *Nat. Commun.* **15**, 8268 (2024).
276. Segal, J. B. *et al.* Assessing heterogeneity of treatment effect in real-world data. *Ann. Intern. Med.* **176**, 536–544 (2023).
277. Stuart, E. A., Bradshaw, C. P. & Leaf, P. J. Assessing the generalizability of randomized

trial results to target populations. *Prev. Sci.* **16**, 475–485 (2015).

278. Trosten, D. J., Strauman, A. S., Kampffmeyer, M. & Jenssen, R. Recurrent deep divergence-based clustering for simultaneous feature learning and clustering of variable length time series. in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2019). doi:10.1109/icassp.2019.8682365.
279. Ienco, D. & Interdonato, R. Deep semi-supervised clustering for multi-variate time-series. *Neurocomputing* **516**, 36–47 (2023).
280. Welcome to the UCR Time Series Classification/Clustering Page. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.
281. UCI Machine Learning Repository. https://archive.ics.uci.edu/.
282. Time Series Classification Website. https://www.timeseriesclassification.com/dataset.php.
283. Qin, Y., Schaar, M. & Lee, C. T-phenotype: Discovering phenotypes of predictive temporal patterns in disease progression. *AISTATS* **206**, 3466–3492 (2023).
284. Aguiar, H., Santos, M., Watkinson, P. & Zhu, T. Learning of Cluster-based Feature Importance for Electronic Health Record Time-series. in *Proceedings of the 39th International Conference on Machine Learning* (eds. Chaudhuri, K. et al.) vol. 162 161–179 (PMLR, 17--23 Jul 2022).
285. Wang, D., Ma, X., Schulz, P. E., Jiang, X. & Kim, Y. Clinical outcome-guided deep temporal clustering for disease progression subtyping. *J. Biomed. Inform.* **158**, 104732 (2024).
286. Huang, Y., Axsom, K. M., Lee, J., Subramanian, L. & Zhang, Y. DICE: Deep significance clustering for outcome-aware stratification. *arXiv [cs.LG]* (2021).
287. Huang, Z., Hao, H. & Du, L. Exploring the explainability of time series clustering: A review of methods and practices. in *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining* 1005–1007 (ACM, New York, NY, USA, 2025).
288. Yerbury, L. W., Campello, R. J. G. B., Livingston, G. C., Jr, Goldsworthy, M. & O'Neil, L. Comparing clustering approaches for smart meter time series: Investigating the influence of

dataset properties on performance. *Appl. Energy* **391**, 125811 (2025).

289. Degen, I., Abdallah, Z. S., Reeve, H. W. J. & Brown, K. R. CSTS: A benchmark for the discovery of Correlation Structures in Time Series clustering. *arXiv [cs.LG]* (2025).

290. Bhandari, P. Levels of Measurement. *Scribbr* https://www.scribbr.com/statistics/levels-of-measurement/ (2020).

291. Hu, L., Jiang, M., Dong, J., Liu, X. & He, Z. Interpretable Clustering: A Survey. *arXiv [cs.LG]* (2024) doi:10.48550/ARXIV.2409.00743.