Counterfactual Explanations and Predictive Models to Enhance Clinical Decision-Making in Schizophrenia using Digital Phenotyping

Juan Sebastián Cañas[1], Francisco Gómez[1], and Omar Costilla-Reyes*[2]

[1]Universidad Nacional de Colombia
[2]CSAIL MIT

June 6, 2023

Abstract

Clinical practice in psychiatry is burdened with the increased demand for healthcare services and the scarce resources available. New paradigms of health data powered with machine learning techniques could open the possibility to improve clinical workflow in critical stages of clinical assessment and treatment in psychiatry. In this work, we propose a machine learning system capable of predicting, detecting, and explaining individual changes in symptoms of patients with Schizophrenia by using behavioral digital phenotyping data. We forecast symptoms of patients with an error rate below 10%. The system detects decreases in symptoms using changepoint algorithms and uses counterfactual explanations as a recourse in a simulated continuous monitoring scenario in healthcare. Overall, this study offers valuable insights into the performance and potential of counterfactual explanations, predictive models, and change-point detection within a simulated clinical workflow. These findings lay the foundation for further research to explore additional facets of the workflow, aiming to enhance its effectiveness and applicability in real-world healthcare settings. By leveraging these components, the goal is to develop an actionable, interpretable, and trustworthy integrative decision support system that combines real-time clinical assessments with sensor-based inputs.

1 Introduction

Schizophrenia spectrum disorders (SSDs) are a family of serious mental illnesses (SMI) affecting approximately 24 million people worldwide. There are still challenges in providing healthcare access and treatment. People with SSDs that do not receive specialist mental health care are close to 70% and just 30% of them fully recover World Health Organization (2022). Current clinical practices, such as conventional face-to-face assessments, are inefficient in detecting early observable behavioral precursors of schizophrenia, cannot scale and are not optimal to detect dynamic behavioral changes, as the nature of such diseases. Consequently, this leads to intervention at late stages of relevant clinical events Ben-Zeev et al. (2017).

Mobile computing technology has opened the possibility for the study of behavoral health conditions in a new way. The act of measurement no longer needs to be confined to clinics or research laboratories. Instead, it can be carried out in real-world settings. As a consequence, a new source of clinical data can be made available to turn it into biomedical knowledge and clinical insights Sim (2019). Given that these digital fingerprints reflect lived experiences of people in their natural environments, with the granular temporal resolution, it might be possible to leverage them to develop precise and temporally dynamic digital disease phenotypes and markers to diagnose and treat psychiatric illnesses and others Perez-Pozuelo et al. (2021). In this research, we explore the area of *digital phenotyping*, understood as the in-situ quantification of the individual-level

*corresponding author: costilla@mit.edu. Working paper.

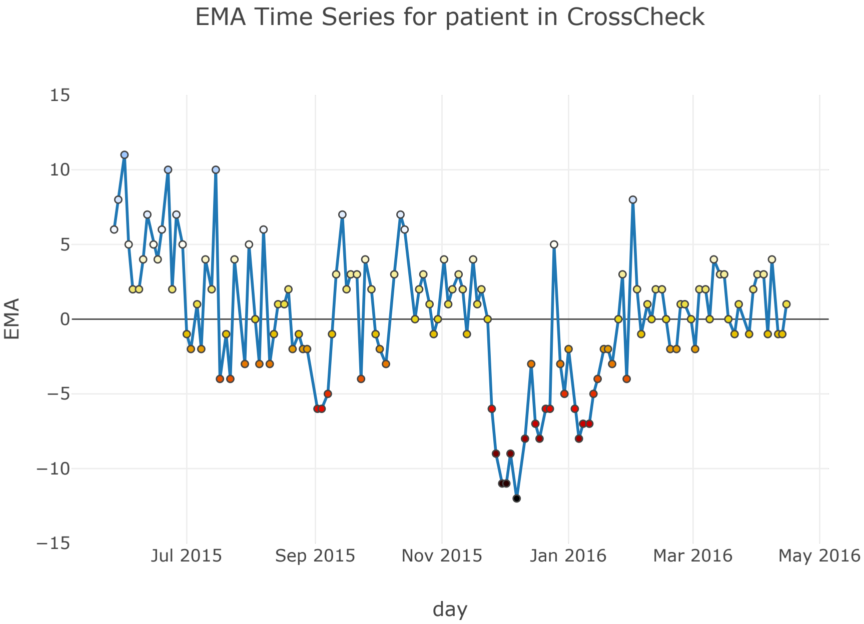

Figure 1: Here we show the trajectory of symptoms report for one user in the study. Starting from November 2015, there is a noticeable *decrease* in the reported symptoms. Subsequently, the patient's symptoms gradually improve over time. Our focus lies in identifying the clinical interventions preceding this textit decreases in symptoms

human phenotype using data from digital devices Mohr et al. (2017); Onnela and Rauch (2016) for the study of schizophrenia.

In conventional clinical practice, healthcare providers engage in a process of gathering patient information, employing their clinical expertise to make informed decisions, and subsequently documenting their findings. In contrast, *integrative decision support systems* (IDSS), can actively request clinically relevant information or gather the data, show the results to clinicians, and support decisions that clinicians still need to make Yu et al. (2018). The goal of this research is to create actionable clinical systems capable of exploiting newly available data from digital phenotyping empowered with data science tools that could have an impact on the clinical task, such as early detection and intervention in mental healthcare Russ et al. (2019).

Driven by the challenge of intervening in the advanced stages of SSDs and recognizing the potential benefits of mobile computing in clinical practice, we have formulated the following research question: *How can we effectively characterize the transitions between different symptom states for patients diagnosed with Schizophrenia using Digital Phenotyping?*

Figure 1 shows the temporal fluctuations in behavioral data exhibited by patients. This graph illustrates the progress of a patient with SSDs who demonstrates high adherence to the treatment. Notably, following a period of relatively stable symptoms from July 2015 to November 2015, the patient reports a decrease in symptoms. In our study, we define a "decrease" as a period where, after experiencing positive symptom levels for some time, there is a subsequent diminishing of symptoms. It is crucial to differentiate a decrease from the clinical event known as a "relapse." A relapse encompasses various events such as psychiatric hospitalization, a significant increase in psychiatric care, the presence of suicidal ideation with clinical relevance, self-injury, or violent behavior resulting in harm to oneself or others Adler et al. (2020); Buck et al. (2019). Relapse can be identified through assessments or electronic medical records, whereas decreases are self-reported reductions in symptoms. We hypothesize that digital phenotyping can be utilized to detect and analyze these decreases in symptoms.

Beyond benchmarks in predictive models, which have been the traditional focus of research in this field, we aim to shift our attention toward actionable models that can make a meaningful impact on clinical practice.

The importance of behavioral feature analysis has been widely recognized in clinical practice Tonekaboni et al. (2019). In this study, we employ counterfactual explanations Verma et al. (2020), a local interpretability method Stiglic et al. (2020), as a tool in IDSS for analyzing individual patient predictive models. Finally, we generate the feature importance of the clinical alert by evaluating the impact of *decreases* in the clinical output.

The clinical relevance of this research is to present a machine learning system as a potential tool for early intervention in patients. This system changes the conventional clinical practice towards a broader support system for physicians in decision-making. Concretely, our contributions are 1) Integration of a machine learning system in healthcare composed of prediction, detection, and explanation; 2) Using counterfactuals to inform which and when to intervene with relevant features in temporal settings with digital phenotyping.

The prediction component of our machine learning system focuses on leveraging digital phenotyping data to develop models that can accurately forecast future symptom states for individual patients. These predictive models use historical patient data from various sources, such as self-reports and sensor data, to capture the temporal patterns and trends in patients symptoms. By training these models on a large dataset, we can identify the factors that contribute to symptom changes and make reliable predictions about future symptom trajectories.

In addition to prediction, the detection component of our system is designed to identify clinically relevant events, specifically decreases in symptoms, which may indicate a need for intervention. We employ change point detection algorithms to analyze the temporal patterns in symptom data and identify significant shifts or deviations from the expected behavior. This enables us to detect periods where patients experience a decrease in symptoms after a period of relatively stable or desired symptom states. By promptly detecting these symptom decreases, clinicians can intervene and provide appropriate support to prevent further health deterioration.

The interpretability component of our machine learning system, lies in *explanations*, which utilizes counterfactuals research. Counterfactual explanations provide insights into the causal relationships between features and outcomes by generating hypothetical scenarios in which a specific feature is modified while keeping other variables constant. By applying counterfactual explanations to our predictive models, we can identify the features that are most influential in driving symptom changes and determine when and which interventions may be effective. This interpretability component empowers clinicians to understand the underlying mechanisms behind the predictions and make informed decisions regarding patient care.

By integrating prediction, detection, and explanation components into a unified clinical system, we aim to provide clinicians with actionable insights and decision support. This system has the potential to transform conventional clinical practice by incorporating data-driven approaches and enabling proactive interventions for patients with schizophrenia. By leveraging digital phenotyping data and state-of-the-art machine learning techniques, we can enhance the understanding and management of symptoms in real-world settings, leading to improved patient outcomes and personalized care.

2 Related Work

Digital Phenotyping aims to extract relevant clinical features from sensing data Mohr et al. (2017). In this line of work, a stability index measurement was proposed by He-Yueya et al. (2020) as an interpretable and clinically relevant feature for intervention. Tseng et al. (2020) presented a system that leverages human rhythms as features for symptom prediction using multi-task learning. They demonstrate the potential of using human rhythms for early detection and intervention without burdening patients or clinicians. The paper evaluates linear and non-linear models, highlighting the trade-off between prediction accuracy and interpretability.

The CrossCheck Project Ben-Zeev et al. (2017); Wang et al. (2016) conducted a clinical trial aimed at developing sensing, inference, and analysis techniques for detecting changes, relapse, prediction, and early intervention in patients with schizophrenia based on observational behavioral precursors. The project's clinical relevance lies in automatically alerting clinicians promptly to prevent or reduce the severity of relapse in patients with mid-symptoms of severe mental disorders. The CrossCheck symptom prediction system utilizes passive sensing and self-report data from phones to track schizophrenia patients' symptoms. Wang et al. (2020b,a) utilized principal components as behavioral patterns and found that applying them reduces feature dimensionality and generates more useful features for prediction. Adler et al. (2020) developed an encoder-decoder neural network-based anomaly detection model using passive sensing data to predict behavioral anomalies indicative of early warning signs for psychotic relapse. They conducted a post hoc analysis using clinical notes to interpret the detected anomalies within the context of severe mental disorders. While insightful, these approaches rely on clinical information to generate interpretations, making it challenging to adapt in real-world monitoring scenarios.

Stachl et al. (2020) explored the prediction of individuals' personality dimensions using smartphone behavioral data. The study investigated the predictive power of various classes of behavioral information, including communication, social behavior, music consumption, app usage, mobility, overall phone activity, and day- and night-time activity. However, the paper also acknowledges the potential privacy implications and dangers associated with the widespread collection and modeling of smartphone behavioral data and there is no actionable information in the study.

Beyond the CrossCheck project, we identified three lines of work related to the explanation of clinical time series. Tonekaboni et al. (2020) focused on explaining model predictions by identifying the relevant observations over time. They developed a generative model to capture the time series dynamics and determined feature importance by quantifying the shift in the predictive distribution over time. Hardt et al. (2020) used gradient methods and a dynamical system to capture predicted risk increases in clinical time series. Crabbé and van der Schaar (2021) explored computer vision methods to transform time series into images and applied perturbation-based detection methods to identify salient features. These three papers evaluated their tools using classical clinical data, such as intensive care unit data, to simulate real hospital monitoring scenarios.

In our paper, we employ the cumulative sum (CUSUM) algorithm for detecting decreases in self-reports and weekly predictions. The CUSUM algorithm is a sequential analysis technique originally introduced by Page (1954) and has since found applications in various domains, including finance, industrial monitoring, public health monitoring, and clinical indicator analysis Sibanda and Sibanda (2007); Suman and Prajapati (2018); Wohl (1977).

Counterfactual explanations aim to provide insights into the decision-making process of predictive models by identifying the minimal set of changes required to alter the model's output. They allow exploration of hypothetical scenarios and answer questions such as "What if?" For a given input $\mathbf{x}$ with a model output of y, counterfactual explanations provide information on how changing $\mathbf{x}$ to $\mathbf{x}_{cf}$ would affect the model's output. The Counterfactual Explanations (CFE) algorithm Verma et al. (2020) is a local interpretability method that enables specific queries and recourse for predictive models Molnar (2020). By following the formulations and implementations in Wachter et al. (2017) and Mothilal et al. (2020), we search for counterfactuals that possess key interpretability properties.

By using the CFE algorithm or similar approaches, researchers and practitioners can search for counterfactual explanations that possess important interpretability properties. These properties could include attributes like proximity to the original input, feasibility, and understandability, among others. By providing information on how changing the input would affect the model's output, counterfactual explanations enhance our understanding of predictive models and enable us to gain insights into their decision-making process.

3 Methods

We simulate a realistic setting of continuous clinical monitoring using digital phenotyping, following the clinical workflow presented in Wang et al. (2018). The system generates reports on adherence and trends, enabling clinicians to proactively reach out to patients when the system predicts an increased risk. To further enhance interpretability, we propose an additional step of local interpretability that identifies the location of potentially relevant clinical events and the features that should be intervened upon.

System overview. In the depicted setting shown in Figure 2, we expect to observe a patient's data for a duration longer than one month. Subsequently, we generate symptom predictions based on the available historical data (Step 1 in Figure 2). For each weekly prediction, we employ a change point algorithm that identifies potential decreases in symptoms by analyzing the time series comprising previous self-reports and predicted self-reports (Step 2 in Figure 2). If a clinical alert in the form of a change point is detected, we

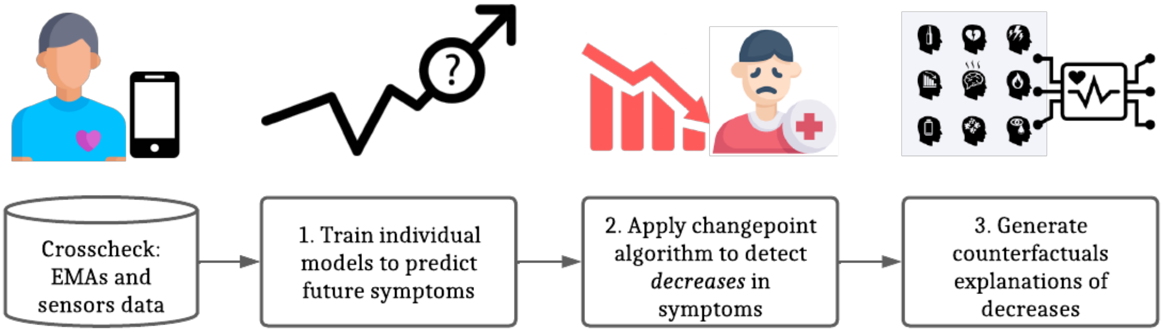
Crosscheck: EMAs and sensors data
1. Train individual models to predict future symptoms
2. Apply changepoint algorithm to detect *decreases* in symptoms
3. Generate counterfactuals explanations of decreases

Figure 2: The machine learning system consists of prediction, detection, and explanation components. This pipeline is applied to individualized models using more than one month of digital phenotyping data

apply counterfactual explanations to the weekly predictions (Step 3 in Figure 2). This workflow can be viewed as a predictive clinical tool that supports decision-making in a continuous monitoring setting.

Problem Statement. Let $X \in \mathbb{R}^{d \times n}$ denote a multi-variate time series data where d is the number of features with n observations over time and $\mathbf{y} \in \mathbb{R}^n$ a univariate time series of size n. At each time step $t = 1, 2, ..., n$ a real value vector, $\mathbf{x}_t = (x_t^1, x_t^2, ..., x_t^d)^\intercal \in \mathbb{R}^d$ and the scalar $y_t \in \mathbb{R}$ is provided. We consider the problem of explaining an alert in a time segment $i : j$ where $i < n < j$ include both historical values (where $t \leq n$) and *predictive* (where $t > n$) of the output y_t, that is, an alert generated from a temporal dynamic of the vector $(y_i, ..., y_n, ..., y_j)$. We are interested in actionable information from the set of features X to inform physicians about possible explanations of the provided alert.

Dataset. The CrossCheck dataset was collected from a psychiatric hospital in New York. The patients were 18 years old or older, met DSM criteria for schizophrenia, schizoaffective disorder, or psychosis, and had a psychiatric hospitalization, daytime psychiatric hospitalization, outpatient crisis management, or short-term psychiatric hospital emergency room visits within 12 months before the study entry Wang et al. (2016). In the dataset, 61 out of 75 participants completed the full year-long study.
Ecological Momentary Assessments (EMAs) or self-reports are clinically validated questions that capture various dynamic dimensions of mental health and functioning in individuals with schizophrenia. The study employs EMAs comprising 10 questions, as presented in Table 1. A higher score on the positive questions indicates better outcomes, while higher scores on the negative item questions suggest worse outcomes. The questions are formulated as concise one-sentence queries, and respondents choose from multiple-choice answers ranging from 0 to 3 Wang et al. (2016).
The second kind of data is passive data from raw sensors collected without the active intervention of patients. The behavioral sensing features are composed of sensors about activity, speech, conversation, calls and SMS, sleep location, phone and app usage, and ambient environment. The dataset is composed of aggregated values of sensors for each day at different time epochs: morning (6 am to 12 pm), afternoon (12 pm to 6 pm), evening (6 pm to 12 am), and night (12 am to 6 am). A more detailed description of this dataset can be found in Wang et al. (2016, 2020b).

Preprocessing EMAs. In our context, we define a *block* as a segment of data that is continuous in time. To construct a block, we consider two parameters: the minimum size of a block and the maximum distance between each data point. For the minimum size of a block, we use 60 days, which corresponds to 15 data points. This ensures that we have enough data to build predictive models and make reliable predictions.
Regarding the maximum distance between each data point, we set a limit of 15 days (equivalent to 6 data points) based on previous preprocessing techniques described in Wang et al. Wang et al. (2016). This means that we consider a gap of up to 15 days between self-report data points within a block.

Category	Question
Positive	1. Have you been feeling CALM? 2. Have you been SOCIAL? 3. Have you been SLEEPING well? 4. Have you been HOPEFUL about the future? 5. Have you been able to THINK clearly?
Negative	6. Have you been worried about people trying to HARM you? 7. Have you been bothered by VOICES? 8. Have you been feeling STRESSED? 9. Have you been SEEING THINGS other people can not see? 10. Have you been DEPRESSED?

Table 1: Questionnaire related to indicators of mental health used in the CrossCheck project. Options: 0—not at all; 1—a little; 2—moderately; 3—extremely

Furthermore, previous studies Adler et al. (2020); Wang et al. (2020b) have shown that there are behavioral patterns and signs of fluctuations close to the 30-day mark, which could be indicative of an impending relapse. Therefore, it is more suitable to think of the *weekly* continuous monitoring as a monitoring period consisting of 3 data points within the block range. By defining blocks in this way, we can effectively analyze and interpret the temporal patterns and fluctuations in the data for predictive modeling and monitoring purposes. In our filtering approach, we employ a variance filter on the EMAs time series. This helps us identify self-reports that do not exhibit dynamic behavior or fluctuations over time. Since our research is specifically interested in analyzing fluctuations in symptoms, a patient whose symptoms remain constant throughout would not provide relevant insights for our research question. By applying this filter, we aim to include patients whose self-reports reflect meaningful variations in their symptoms, thereby ensuring that our machine-learning and counterfactual models capture the dynamic nature of mental health conditions and provide more accurate predictions and insights.

Predictive Modeling. In the first part of the framework in Figure 2 we use individual forecast EMA sum scores using past values of each 10 past EMAs or sensors. An individual model is a fully personalized model, which uses data only from the subject to train the model. Several studies in the CrossCheck project have shown that the individual models outperform population-level ones He-Yueya et al. (2020); Wang et al. (2016). Formally, we consider a multivariate time series $X \in \mathbb{R}^{d \times n}$ and $Y \in \mathbb{R}^{n}$ a univariate time series, where d is the number of features with n observations over time. We are interested in the multi-step ahead point forecasting problem that involves producing an estimate of the H future values $y_{n+1}, y_{n+2}, ..., y_{n+H}$, where $H > 1$ is the forecast horizon using a data drive prediction model composed by a predictive function f and a set of features X.

Hyper-parameter tuning. We search hyperparameters for several predictors. Motivated by the result of Tseng et al. (2020) where they found that nonlinear models have higher prediction accuracy than linear models. We explore the trade-off between linear and nonlinear models as in Zihni et al. (2020) Stachl et al. (2020). In our case, we compare linear and ensemble models to forecast weekly symptoms using digital phenotyping. Specifically, we compare linear models such as Lasso and Elastic Net and ensemble methods such as Random Forest and Gradient Boosting. After obtaining the optimal hyperparameters, we select the best model and use it in the rest of the system.
For each block, the data were ordered and split into training and test sets with a corresponding 80% and 20%. This approach allowed us to test for optimal model settings while ensuring a strict separation of training and test data, especially in the data-dependent case. Then the models were tuned using 5-fold time series forward cross-validation. Table 4 provides a summary of the tuned hyperparameters for each model. We used a grid search approach for the tuning of hyperparameters in all models Pedregosa et al. (2011).

Model training. We used the best-tuned predictors in the simulated clinical workflow. In this part, we test performance using the time series forward (rolling origin) Cross-Validation approach Bergmeir and Benítez (2012). In this case, for each block, the minimum date of the test set is always bigger than the maximum date of the training set. In this order, the size of the training set always increases and the size of the test set is always the size of weekly predictions. This is coherent with a real scenario of monitoring and the time-series properties. For testing, we use the Mean Absolute Error (MAE) that corresponds to the expected value of the absolute error loss: $\text{MAE}(y, \hat{y}) = \frac{1}{H} \sum_{k=n+1}^{n+H} |y_k - \hat{y_k}|$.

Change point Detection. Here we describe the change point detection applied to the symptoms time series, as depicted in the second part of the framework shown in Figure 2. A change point refers to an abrupt shift in a time series, indicating a significant change in the underlying dynamics Basseville et al. (1993). The goal of change point detection is to identify the times within a given time horizon when such changes occur. The strength of the CUSUM approach Page (1954) lies in its simplicity and visual clarity. The algorithm operates online, making it suitable for early detection. We begin by defining a segment of the study, denoted as $\mathbf{y}i : j = (y_i, yi + 1, ..., y_{j-1}, y_j)^\intercal$, which represents a portion of the time series. A reference point, typically chosen as the middle point of the segment, divides it into two subsegments. We calculate the means of each subsegment, denoted as μ_a and μ_b, respectively. We then compute the mean difference, $\mu = \frac{\mu_a - \mu_b}{2}$. Next, we apply the CUSUM function to find the argument that maximizes the cumulative sum of the difference between each data point in the segment of study and μ. Formally, we aim to solve the following optimization problem:

$$\arg\max_k c(y[i : j]) = \sum_{k=i+1}^{N} \|y[k] - \mu\| \text{ for } N = i + 1, ..., j. \tag{1}$$

The algorithm used in our approach estimates the means before and after a potential change point. It iteratively searches for the change point by maximizing the cumulative sum value until either a stable change point is found in the segment or the maximum number of iterations is reached Truong et al. (2020).
To adapt this algorithm to the continuous monitoring setting, we employ a sliding window approach. For each week, we perform CUSUM searches within a window of interest. This window consists of a historical window, which includes the previous 12 data points, a scan window, which includes the next 6 data points, and a step size of 1 data point. Following the search process described above, we evaluate the potential change point using a Gaussian distribution as the underlying model. We apply the log-likelihood ratio test to determine if there is a statistically significant change in the mean of the time series. The null hypothesis assumes no change in the mean, while the alternative hypothesis suggests a change point with two distinct means Jiang et al. (2022). In the continuous monitoring setting, we collect all the potential change points detected within each week and select the most recent change point identified across the all-time series. If a change point is detected within a 2-week window (consisting of 1 week of historical data and 1 week of predictions), we trigger a clinical alert. This approach enables the timely detection of significant changes in the time series, allowing for proactive interventions when necessary.

Counterfactual Explanations. The third part of the framework in Figure 2 focuses on explaining the alerts generated by the change point algorithm using the predictive model. To achieve this, we employ the counterfactual explanations algorithm (CFE) Verma et al. (2020), which aims to identify the minimal set of changes necessary to alter the model's output. CFE operates by searching for hypothetical scenarios, asking questions like: "Given that the model f produces output y for input $\mathbf{x}$, what would be the output if $\mathbf{x}$ were changed to $\mathbf{x}_{cf}$?" In other words, it determines the changes needed in $\mathbf{x}$ to achieve a desired change in the model's output. CFE serves as a local interpretability method, enabling specific queries about the predictive model Molnar (2020). We follow the formulation and implementations presented in Wachter et al. (2017) and Mothilal et al. (2020), seeking counterfactuals with certain properties. These properties include being *proximal* to highlight the local decision logic of the predictor, *sparse* to emphasize a limited set of features, *diverse* to showcase different ways of achieving the same outcome, and *feasible*, meaning the changes in a counterfactual example should be within the possible range of each feature.
Consider a d-dimensional vector $\mathbf{x}$ representing a specific instance at a given time t. We have a machine learning model f that provides predictions $f(\mathbf{x})$, where the predicted values fall within a certain domain.

Within this domain, some subsets correspond to desired outcomes, while others correspond to undesired outcomes. The goal of counterfactual explanation is to find a counterfactual instance $\mathbf{x}cf$ that satisfies the following conditions: (1) $f(\mathbf{x}cf)$ falls into the desired outcome subset, (2) $\mathbf{x}cf$ is close to $\mathbf{x}$, and (3) $\mathbf{x}cf$ is both feasible and plausible.

To obtain a counterfactual explanation, we formulate an optimization problem:

$$\begin{aligned} \arg\min_{\mathbf{x}cf} \quad & \text{dist}(\mathbf{x}, \mathbf{x}cf) \\ \text{s.t.} \quad & f(\mathbf{x}cf) \text{ falls within the desired outcome range} \\ & \mathbf{x}cf \in P \\ & \mathbf{x}cf \in F(\mathbf{x}) \end{aligned} \tag{2}$$

Here, dist represents a distance function that measures the proximity between $\mathbf{x}$ and $\mathbf{x}cf$. The counterfactual explanation operates within the domain space $\mathcal{X} \subset \mathbb{R}^d$ and is subject to plausibility constraints denoted by P and feasibility constraints denoted by $F(\mathbf{x})$. These constraints ensure that the counterfactual instance is both plausible and feasible.

In our specific case, we employ counterfactual explanations using a forecasting model that predicts the EMA sum score. This means that $f(\mathbf{x})$ represents the predicted EMA sum score based on the features in $\mathbf{x}$. By generating counterfactual instances that satisfy the conditions outlined above, we can provide explanations for instances where the predicted EMA sum score falls into the undesired outcome subset. These counterfactual explanations help us understand the changes necessary in the input features to achieve a desired outcome, making them valuable for decision-making and intervention planning.

4 Results

In Figure 3 (left), a complete execution of our simulated clinical workflow for a specific patient is depicted. The system initiates by monitoring relevant outcomes in the form of time series. Predictions are generated (depicted as red triangles). Subsequently, the change point algorithm is applied to the merged historical and weekly predictive time series. If a change point is detected, the CFE algorithm generates a hypothetical situation where the patient exhibits improved symptoms (represented by green squares). Clinicians can then intervene in the features that differ between the predictions and the counterfactuals.

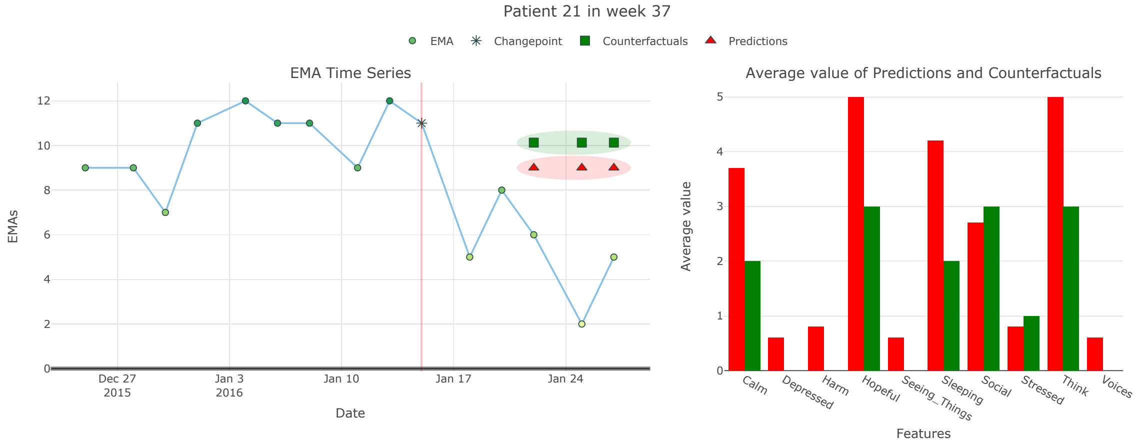

Figure 3: Results of continuous symptom monitoring and counterfactual explanation. Historical data (left) shows symptom values of 12 to 7 before January 17. In under a week, decreasing symptoms were detected and counterfactuals generated (right). The system identifies symptom changes, determines modifications needed, and helps prevent adverse health events.

In Figure 3 (right), the alert generated highlights which features change and the magnitude of their changes. This information enables our system to identify the location of an alert, the features involved, and the extent of changes required to achieve the desired outcome.

features	method	validity	redundancy	sparsity	proximity
EMAs	**genetic**	0.9(0.10)	**0.19(0.16)**	0.75(0.16)	**0.45(0.46)**
	kdtree	1.0(0.0)	0.15(0.18)	0.79(0.2)	0.34(0.43)
	random	1.0(0.0)	0.15(0.08)	**0.82(0.04)**	0.18(0.12)
Sensors	**genetic**	0.85(0.11)	**0.67(0.31)**	0.22(0.14)	**0.53(0.57)**
	kdtree	0.91(0.09)	0.67(0.32)	0.21(0.14)	0.52(0.6)
	random	1.0(0.0)	0.52(0.23)	**0.4(0.11)**	0.01(0.01)
Sensors +EMAs	**genetic**	0.9(0.13)	0.6(0.33)	0.25(0.13)	**1.72(11.41)**
	kdtree	0.87(0.11)	**0.62(0.32)**	0.24(0.13)	1.71(11.4)
	random	1.0(0.0)	0.45(0.25)	**0.44(0.1)**	0.01(0.01)

Table 2: Evaluation of counterfactual explanations methods

After preprocessing the dataset, we obtained a total of 44 patients and 51 blocks. The minimum block size is 26 data points (equivalent to approximately two months and one week), while the maximum block size is 165 data points. On average, each block contains 91 data points.

Definitions. *Validity* refers to the fraction of examples that are actually counterfactuals, meaning they result in a different outcome than the original input. *Proximity* represents the average distance between a counterfactual example and the original input, calculated feature-wise. Proximity is evaluated by taking the mean of the distances for a set of examples. *Sparsity* captures the number of features that are different between the original input and a generated counterfactual. It is defined as the number of changes between the two inputs. *Diversity* can be evaluated by measuring the feature-wise distances between each pair of counterfactual examples. The mean distance between pairs represents the diversity of the set. Separate diversity metrics are computed for categorical and continuous features. *Redundancy* refers to the duplicated or highly similar information across multiple counterfactual explanations. High redundancy means the method is proposing the same or nearly identical explanations repeatedly, indicating limited diversity.

The research results presented in Table 2 compare the performance of different counterfactual explanation methods across three feature sets: EMAs, Sensors, and Sensors+EMAs. The evaluation is based on four key metrics: validity, redundancy, sparsity, and proximity. The counterfactual explanation methods employed are genetic, kdtree, and random.
The results show that all methods consistently achieved high validity scores, indicating their effectiveness in producing meaningful counterfactual explanations. The genetic method consistently outperforms the other methods in terms of redundancy for both the EMAs and Sensors feature sets. This suggests that the genetic method generates a more diverse set of counterfactuals by exploring different regions of the feature space. The random method consistently exhibited the highest sparsity across all feature sets, implying that it introduces a larger number of changes compared to the other methods. Lower proximity scores indicate that the counterfactuals are closer to the original input. Interestingly, there is no clear pattern among the methods in terms of proximity scores, as the values vary across feature sets and methods.
Evaluating counterfactual generation methods requires consideration of validity in relation to redundancy, sparsity, proximity, and diversity. A nuanced understanding of performance across metrics can determine the optimal method for developing meaningful counterfactual explanations in a given context and support application to diverse datasets.

4.1 Predictive performance evaluation

In our simulated setting, we used the information from the three previous self-reports to forecast the EMA sum score. Individual predictions were initiated with a minimum of 12 data points and increased weekly. As depicted in Figure 4, all predictors demonstrated comparable performance. These models were trained using 12 data points based on the week of predictions, and the training data consisted of all EMA questions. The results show that there is minimal variation among the models in this setup. The error bars represent 95% bootstrapped confidence intervals. Specifically, the MAE values in the test set were as follows: mean baseline

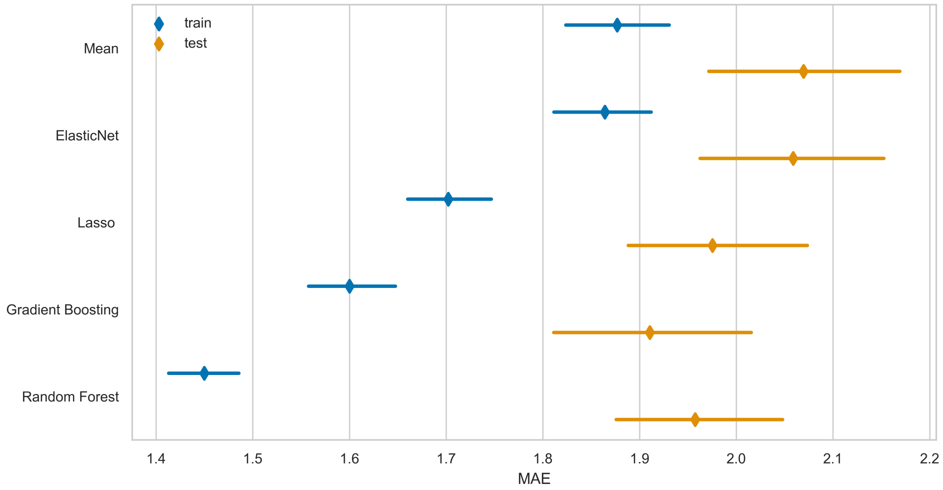

Figure 4: Comparison of Train and Test Error with MAE for Different Models.

- 2.070, Elastic Net - 2.059, Lasso - 1.975, Random Forest - 1.958, and GBRT - 1.911. Although the GBRT model exhibited the highest performance, the difference in performance among the models was marginal. It is worth noting that the nonlinear models consistently outperformed the linear prediction performance on average. Therefore, we selected the GBRT model for all subsequent predictions in this paper.

After conducting the cross-validation process, we successfully determined the optimal hyperparameters for each predictor. The Lasso model yielded an alpha value of 0.74. For the Elastic Net model, we obtained an alpha of 10 and an L1 ratio of 0.14. The Random Forest model was configured with a maximum tree depth of 5, consisting of 10 trees. We set the minimum number of samples required to split an internal node to 2, and the minimum number of samples required at a leaf node to 3. Lastly, the Gradient Boosting model utilized a Huber loss function, a learning rate of 0.05, 10 boosting stages, a maximum depth of estimators set to 3, and minimum numbers of samples required at a leaf node and to split an internal node set to 5.

Figure 5 presents a comparison of feature performance for the GBRT-Huber model. The features evaluated include the historical EMA sum score, all EMA responses, and sensor data. The error bars in the figure represent bootstrapped confidence intervals at a 95% level.

As shown in Figure 5, there are no significant differences in performance among the different features. This finding suggests that it is feasible to use raw sensor data as forecasting features, opening the possibility of leveraging sensor information for behavior prediction.

By utilizing the GBRT model and conducting weekly predictions, we obtained the following mean MAE values: 1.911 for all past EMAs, 1.908 for the past EMA sum score, and 2.006 for the sensor data. Although all past EMAs and the past EMA sum score demonstrate slightly better performance, we select all past EMAs as the feature set for our final model. This choice is based on its clearer clinical interpretation and relevance to the forecasting task.

These results are encouraging, as they suggest that raw sensor data can be leveraged for forecasting behavior and support the application of CFE techniques to sensor data.

4.2 Change points Detection of predicted symptoms

Table 3 presents an evaluation of different changepoint algorithms based on their recall and F1-score, with a 5-day delay.

The Baseline Zero algorithm achieves a recall of 0.53 and the highest F1-score of 0.66. It strikes a good balance between true positives and false positives, leading to a high F1-score. However, its relatively lower recall suggests that it may miss some true positive change points. This algorithm serves as a baseline for comparison.

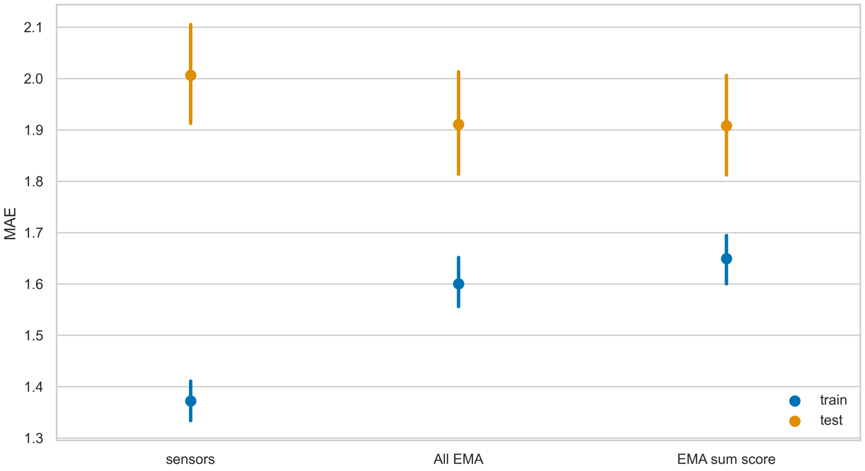

Figure 5: Features performance for the GBRT-Huber model. There are no big differences between different features. This opens the possibility of using raw sensors as forecasting features. The error bars indicate 95% bootstrapped confidence intervals.

Changepoint Algorithm	**Recall**	**F1-Score**
Baseline Zero	0.53	**0.66**
CUSUM	0.73	0.52
CUSUM sliding window	0.61	0.56
Bayesian Online	**0.77**	0.58
Robust Detection	0.64	0.51

Table 3: Evaluation of Changepoint Algorithms using recall and F1 score using 5 days of delay. The best algorithm in terms of recall is the CUSUM with a sliding window. The baseline zero algorithm is the best in terms of F1 score, this could be explained by many false-negative reports. This implies a risk in terms of alarm fatigue.

The CUSUM algorithm exhibits a higher recall of 0.73 compared to Baseline Zero. However, its F1-score is lower at 0.52. This indicates that CUSUM detects more true positive changepoints but also produces more false positives, resulting in a lower overall F1-score.

The CUSUM algorithm with sliding window achieves a recall of 0.61 and an F1-score of 0.56. Although its recall is slightly lower than the original CUSUM algorithm, it manages to improve the F1-score. This suggests a better balance between true positives and false positives compared to the standard CUSUM algorithm.

The Bayesian Online algorithm demonstrates the highest recall of 0.77 among all the evaluated algorithms. However, its F1-score is 0.58, indicating that it also produces a significant number of false positives. Despite this, it may be suitable for applications where high recall is crucial.

The Robust Detection algorithm achieves a recall of 0.64 and an F1-score of 0.51. It exhibits a relatively higher recall compared to the CUSUM algorithm but has a lower F1-score. This implies that it detects more true positive changepoints but also generates more false positives, resulting in a lower overall F1-score.

In summary, the evaluation of these changepoint algorithms reveals trade-offs between recall and F1-score. The CUSUM algorithm with a sliding window strikes a balance between the two metrics. The Baseline Zero algorithm achieves the highest F1-score but has a lower recall, indicating a risk of false-negative reports. On the other hand, the Bayesian Online algorithm has the highest recall but also a notable number of false positives. The choice of algorithm will depend on the specific application and the relative importance of recall and precision in the clinical context.

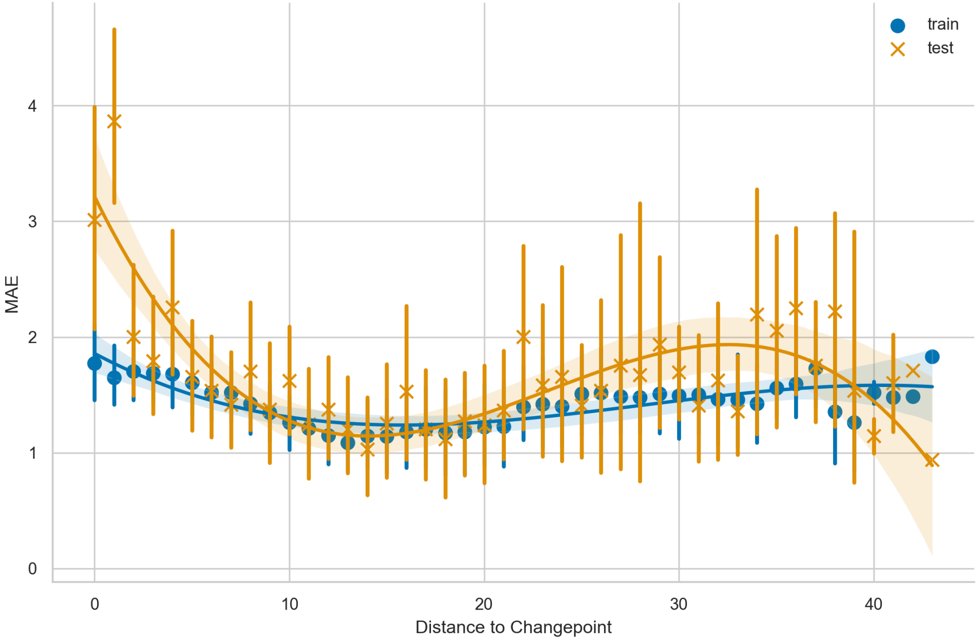

Figure 6: Relationship between distance to a change point and forecasting error. The plot illustrates that as data points approach the change point, the errors in forecasting increase. This observation is likely attributed to the distribution shift occurring in the training data.

We have identified a total of 33 change points distributed across 22 blocks. However, during our analysis, we encountered an unexpected observation regarding the influence of the distance to change points on the algorithm's error. To measure this distance, we consider the number of weeks between a data point and a change point. Figure 6 visually demonstrates that as a data point gets closer to a change point, the forecasting error increases. This phenomenon exemplifies a classic case of distribution shift, which has the potential to impact machine learning systems. Notably, recent research has explored the concept of distribution shift to developing alert systems Tonekaboni et al. (2020). In our future work, we aim to focus on developing prediction models and CFE algorithms that are robust to distribution shifts Rawal et al. (2020). Understanding and addressing the challenges posed by distribution shifts in prediction models and CFE algorithms are crucial for the robustness and reliability of our system. By further investigating and mitigating the effects of distribution shifts, we can enhance the performance and effectiveness of our algorithms in real-world scenarios.

4.3 Counterfactual Explanations of Predicted Symptoms

Figure 7 illustrates the transitions between predictions and counterfactual results, showing the pre-explanation values of each feature and the necessary changes for all patients. Clinicians can use these transitions to identify specific areas requiring attention and potential interventions for each patient.

We generated a total of 33 counterfactual explanations during our simulated clinical workflow. Figure 8 depicts the histogram of EMA sum score values for predictions and counterfactuals. The observed decreases in symptoms generally fall within desirable ranges, which may be attributed to the scarcity of patients with low symptoms in the CrossCheck project and the system's limited ability to detect abrupt changes. Notably, a separate study Adler et al. (2020) identified four relapse patients using clinical labeled data, not EMAs.

In conclusion, our counterfactual explanations provide valuable insights into predicted symptom changes. Although the detected decreases mostly occur within higher symptom ranges, future work should focus on establishing a ground truth to evaluate the accuracy of alerts. These enhancements will improve the reliability and effectiveness of our counterfactual explanation system, supporting clinical decision-making and patient care.

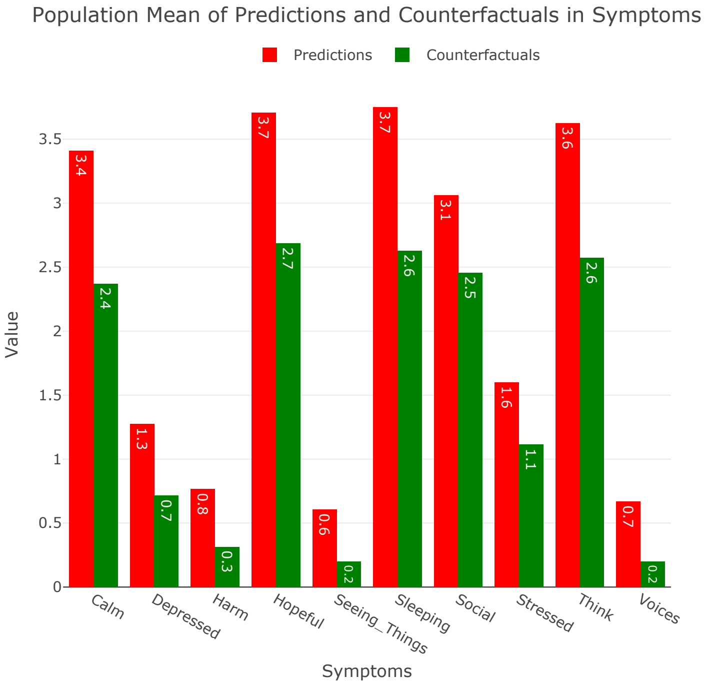

Figure 7: Population values for each EMA question by predictions and counterfactuals.

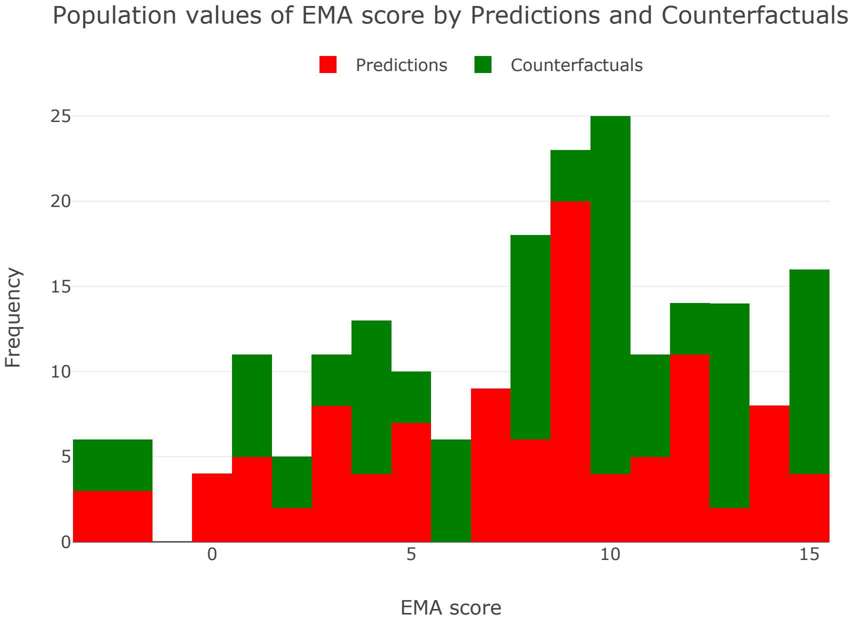

Figure 8: Histogram of values of EMA sum score by predictions and counterfactuals. We can see that detection of decreases still in high symptoms.

5 Conclusions

In conclusion, our study investigated various aspects of a simulated clinical workflow, including counterfactual explanations, predictive performance evaluation, and change point detection of predicted symptoms in schizophrenia patients.

Counterfactual Explanation Methods: The evaluation of different counterfactual explanation methods revealed that all methods achieved high validity scores, indicating their effectiveness in generating meaningful counterfactual explanations. The genetic method consistently outperformed the other methods in terms of redundancy, suggesting that it generated a more diverse set of counterfactuals. The random method exhibited the highest sparsity.

Predictive Performance Evaluation: The comparison of different predictive models showed minimal variation in performance, with marginal differences in MAE values. Nonlinear models consistently outperformed linear models, with the GBRT model exhibiting the highest performance. The choice of the GBRT model for subsequent predictions was based on its slightly better performance and the clinical interpretation of its features.

Change Point Detection: The evaluation of different change point detection algorithms showed trade-offs between recall and F1-score. The CUSUM algorithm with a sliding window achieved a balance between the two metrics, while the Baseline Zero algorithm had the highest F1-score but lower recall. The Bayesian Online algorithm had the highest recall but also more false positives. The choice of algorithm depends on the specific application and the importance of recall and precision in the clinical context.

Overall, our study provides insights into the performance and capabilities of a simulated clinical workflow, highlighting the potential of using counterfactual explanations, interpretability, predictive models, and change point detection for improving patient care and decision-making. Further research can build upon these findings and explore additional aspects of the workflow to enhance its effectiveness and applicability in real-world healthcare settings.

This study has also limitations. First, all models were trained on a small sample of individuals with SSDs. Additionally, assessments of symptoms were reported only on selected days. The small dataset may have contributed to reduced predictive performance, further work should apply this pipeline with clinical datasets with more granular levels (minutes or seconds) and large sample datasets. Additionally, the change point algorithm does not have a ground truth to test the accuracy of alerts. The CFE algorithm assumes that the features in the prediction algorithm are symmetrically related to the outcome. It is important to construct robust prediction models and evaluate the change point and CFE algorithms before deployment in real-world settings.

6 Future directions

In our study, we primarily relied on EMAs and sensor data for predictive modeling and counterfactual analysis. However, there is a wealth of other data sources that can be incorporated to improve the accuracy and robustness of the models. For instance, integrating electronic health records (EHRs), genetic data, social determinants of health, and additional clinical variables can provide a more comprehensive understanding of a patient's health status and enable more accurate predictions. Future work should explore the integration of diverse data sources to create a holistic view of patient's health profiles and enable personalized interventions.

To fully understand the potential impact of our approaches, it is essential to evaluate their clinical utility and effectiveness. Conducting studies to assess the impact of using predictive models and counterfactual explanations in clinical decision-making can provide valuable insights into their value in improving patient outcomes, reducing healthcare costs, and enhancing the quality of care. Future work should focus on conducting rigorous clinical trials and outcome evaluations to quantify the benefits and limitations of the proposed approaches.

As with any technology in healthcare, there are ethical considerations associated with the use of predictive models and counterfactual explanations. Future research should address ethical challenges related to data privacy, consent, algorithmic bias, and the responsible deployment of these technologies in clinical practice. Incorporating ethical frameworks and involving stakeholders, including patients, clinicians, and policymakers, in the design and evaluation process can help mitigate ethical concerns and ensure the responsible use of these tools.

Developing techniques to provide meaningful and transparent explanations for these models would greatly enhance their usability and acceptance in critical domains such as healthcare. Moreover, exploring the integration of counterfactual explanations with other interpretable techniques, such as rule-based models or symbolic reasoning, could provide a more holistic and comprehensive understanding of the model's decision-making process.

Acknowledgment

Research supported by the National Science Foundation (NSF) under Grant No. 1918839. We thank Armando Solar-Lezama and the computer-assisted programming group at CSAIL MIT for fruitful discussions.

References

Daniel A. Adler, Dror Ben-Zeev, Vincent W.S. Tseng, John M. Kane, Rachel Brian, Andrew T. Campbell, Marta Hauser, Emily A. Scherer, and Tanzeem Choudhury. 2020. Predicting early warning signs of psychotic relapse from passive sensing data: An approach using encoder-decoder neural networks. *JMIR mHealth and uHealth* 8, 8 (2020).

Michele Basseville, Igor V Nikiforov, et al. 1993. *Detection of abrupt changes: theory and application*. Vol. 104. prentice Hall Englewood Cliffs.

Dror Ben-Zeev, Rachel Brian, Rui Wang, Weichen Wang, Andrew T. Campbell, Min S.H. Aung, Michael Merrill, Vincent W.S. Tseng, Tanzeem Choudhury, Marta Hauser, John M. Kane, and Emily A. Scherer. 2017. CrossCheck: Integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse. *Psychiatric Rehabilitation Journal* 40, 3 (2017), 266–275.

Christoph Bergmeir and José M Benítez. 2012. On the use of cross-validation for time series predictor evaluation. *Information Sciences* 191 (2012), 192–213.

Benjamin Buck, Emily Scherer, Rachel Brian, Rui Wang, Weichen Wang, Andrew Campbell, Tanzeem Choudhury, Marta Hauser, John M Kane, and Dror Ben-Zeev. 2019. Relationships between smartphone social behavior and relapse in schizophrenia: a preliminary report. *Schizophrenia research* 208 (2019), 167–172.

Jonathan Crabbé and Mihaela van der Schaar. 2021. Explaining Time Series Predictions with Dynamic Masks. *arXiv preprint arXiv:2106.05303* (2021).

Michaela Hardt, Alvin Rajkomar, Gerardo Flores, Andrew Dai, Michael Howell, Greg Corrado, Claire Cui, and Moritz Hardt. 2020. Explaining an increase in predicted risk for clinical alerts. In *Proceedings of the ACM Conference on Health, Inference, and Learning*. 80–89.

Joy He-Yueya, Benjamin Buck, Andrew Campbell, Tanzeem Choudhury, John M. Kane, Dror Ben-Zeev, and Tim Althoff. 2020. Assessing the relationship between routine and schizophrenia symptoms with passively sensed measures of behavioral stability. *npj Schizophrenia* 6, 1 (2020).

Xiaodong Jiang, Sudeep Srivastava, Sourav Chatterjee, Yang Yu, Jeffrey Handler, Peiyi Zhang, Rohan Bopardikar, Dawei Li, Yanjun Lin, Uttam Thakore, Michael Brundage, Ginger Holt, Caner Komurlu, Rakshita Nagalla, Zhichao Wang, Hechao Sun, Peng Gao, Wei Cheung, Jun Gao, Qi Wang, Marius Guerard, Morteza Kazemi, Yulin Chen, Chong Zhou, Sean Lee, Nikolay Laptev, Tihamér Levendovszky, Jake Taylor, Huijun Qian, Jian Zhang, Aida Shoydokova, Trisha Singh, Chengjun Zhu, Zeynep Baz, Christoph Bergmeir, Di Yu, Ahmet Koylan, Kun Jiang, Ploy Temiyasathit, and Emre Yurtbay. 2022. *Kats*.

David C. Mohr, Mi Zhang, and Stephen M. Schueller. 2017. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. *Annual Review of Clinical Psychology* 13, March (2017), 23–47.

Christoph Molnar. 2020. *Interpretable machine learning*. Lulu. com.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 607–617.

Jukka Pekka Onnela and Scott L. Rauch. 2016. Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health. *Neuropsychopharmacology* 41, 7 (2016), 1691–1696.

Ewan S Page. 1954. Continuous inspection schemes. *Biometrika* 41, 1/2 (1954), 100–115.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

Ignacio Perez-Pozuelo, Dimitris Spathis, Emma A.D. Clifton, and Cecilia Mascolo. 2021. Wearables, smartphones, and artificial intelligence for digital phenotyping and health. *Digital Health* 54 (2021), 33–54.

Kaivalya Rawal, Ece Kamar, and Himabindu Lakkaraju. 2020. Algorithmic recourse in the wild: Understanding the impact of data and model shifts. *arXiv preprint arXiv:2012.11788* (2020).

Tom C Russ, Eva Woelbert, Katrina AS Davis, Jonathan D Hafferty, Zina Ibrahim, Becky Inkster, Ann John, William Lee, Margaret Maxwell, Andrew M McIntosh, et al. 2019. How data science can advance mental health research. *Nature human behaviour* 3, 1 (2019), 24–32.

Thabani Sibanda and Nokuthaba Sibanda. 2007. The CUSUM chart method as a tool for continuous monitoring of clinical outcomes using routinely collected data. *BMC medical research methodology* 7, 1 (2007), 1–7.

Ida Sim. 2019. Mobile devices and health. *New England Journal of Medicine* 381, 10 (2019), 956–968.

Clemens Stachl, Quay Au, Ramona Schoedel, Samuel D Gosling, Gabriella M Harari, Daniel Buschek, Sarah Theres Völkel, Tobias Schuwerk, Michelle Oldemeier, Theresa Ullmann, et al. 2020. Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences* 117, 30 (2020), 17680–17687.

Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. 2020. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 5 (2020), e1379.

Gaurav Suman and DeoRaj Prajapati. 2018. Control chart applications in healthcare: a literature review. *International Journal of Metrology and Quality Engineering* 9 (2018), 5.

Sana Tonekaboni, Shalmali Joshi, Kieran Campbell, David Duvenaud, and Anna Goldenberg. 2020. What went wrong and when? instance-wise feature importance for time-series models. *arXiv preprint arXiv:2003.02821* (2020).

Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*. PMLR, 359–380.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. 2020. Selective review of offline change point detection methods. *Signal Processing* 167 (2020), 107299.

Vincent W.S. Tseng, Akane Sano, Dror Ben-Zeev, Rachel Brian, Andrew T. Campbell, Marta Hauser, John M. Kane, Emily A. Scherer, Rui Wang, Weichen Wang, Hongyi Wen, and Tanzeem Choudhury. 2020. Using behavioral rhythms and multi-task learning to predict fine-grained symptoms of schizophrenia. *Scientific Reports* 10, 1 (2020), 1–17.

Sahil Verma, John Dickerson, and Keegan Hines. 2020. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596* (2020).

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

Rui Wang, Min S.H. Aung, Saeed Abdullah, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Michael Merrill, Emily A. Scherer, Vincent W.S. Tseng, and Dror Ben-Zeev. 2016. CrossCheck: Toward passive sensing and detection of mental health changes in people with schizophrenia. *UbiComp 2016 - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2016), 886–897.

Rui Wang, Weichen Wang, Min Hane Aung, Dror Ben-Zeev, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Emily A. Scherer, and Megan Walsh. 2018. Predicting Symptom Trajectories of Schizophrenia Using Mobile Sensing. *GetMobile: Mobile Computing and Communications* 22, 2 (2018), 32–37.

Rui Wang, Weichen Wang, Mikio Obuchi, Emily Scherer, Rachel Brian, Dror Ben-Zeev, Tanzeem Choudhury, John Kane, Martar Hauser, Megan Walsh, and Andrew Campbell. 2020b. On Predicting Relapse in Schizophrenia using Mobile Sensing in a Randomized Control Trial. *18th Annual IEEE International Conference on Pervasive Computing and Communications, PerCom 2020* (2020).

Weichen Wang, Shayan Mirjafari, Gabriella Harari, Dror Ben-Zeev, Rachel Brian, Tanzeem Choudhury, Marta Hauser, John Kane, Kizito Masaba, Subigya Nepal, Akane Sano, Emily Scherer, Vincent Tseng, Rui Wang, Hongyi Wen, Jialing Wu, and Andrew Campbell. 2020a. Social Sensing: Assessing Social Functioning of Patients Living with Schizophrenia using Mobile Phone Sensing. *Conference on Human Factors in Computing Systems - Proceedings* (2020), 1–15.

Herbert Wohl. 1977. The cusum plot: its utility in the analysis of clinical data. *New England Journal of Medicine* 296, 18 (1977), 1044–1045.

World Health Organization. 2022. Schizophrenia. Fact sheet.

Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. 2018. Artificial intelligence in healthcare. *Nature biomedical engineering* 2, 10 (2018), 719–731.

Esra Zihni, Vince Istvan Madai, Michelle Livne, Ivana Galinovic, Ahmed A Khalil, Jochen B Fiebach, and Dietmar Frey. 2020. Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. *Plos one* 15, 4 (2020), e0231166.

A Appendix

Hyperparameter space and description for predictive performance evaluation. After conducting the cross-validation process, we successfully determined the optimal hyperparameters for each predictor. The Lasso model yielded an alpha value of 0.74. For the Elastic Net model, we obtained an alpha of 10 and an L1 ratio of 0.14. The Random Forest model was configured with a maximum tree depth of 5, consisting of 10 trees. We set the minimum number of samples required to split an internal node to 2, and the minimum number of samples required at a leaf node to 3. Lastly, the Gradient Boosting model utilized a Huber loss function, a learning rate of 0.05, 10 boosting stages, a maximum depth of estimators set to 3, and minimum numbers of samples required at a leaf node and to split an internal node set to 5.

Predictor	**Hyperparameter**	**Values**
Lasso	Alpha (Constant that multiplies the L1 term)	$\{\frac{k}{100} : 0 < k < 100, k \in \mathbb{N}\}$
Elastic Net	L1 ratio (The ElasticNet mixing parameter)	$\{\frac{k}{100} : 0 < k < 100, k \in \mathbb{N}\}$
	Constant that multiplies the penalty terms	0.00001 ,0.0001, 0.001, 0.01, 0.1, 0, 0.5, 1, 10, 100
Random Forest	Whether bootstrap samples are used when building trees	True, False
	The maximum depth of the tree	5, 10, 20, 50, 100
	The number of trees in the forest.	5, 10, 50, 100, 500, 1000
	The minimum number of samples required to split an internal node	1, 2, 5
	The minimum number of samples required to be at a leaf node	1, 3, 5
Gradient Boosting	Loss function	Huber, Squared Error
	Learning rate	0.001, 0.01, 0.05, 0.1, 0.2, 1
	The number of boosting stages	5, 10, 50, 100, 500, 1000
	Maximum depth of the individual regression estimators	2, 3, 5, 10
	The minimum number of samples required to be at a leaf node	1, 5, 10, 20
	The minimum number of samples required to split an internal node	2, 5, 10, 20

Table 4: Model hyperparameters and values. These models were tuned before the simulated clinical workflow using time series 5-fold cross-validation using 80% of each model. The 20% of each model was used to evaluate.