Final Degree Project

**Biomedical Engineering Degree**

# Antipsychotic Use in Schizophrenia in Catalonia: A Real-World Comparative Effectiveness Study of Adherence Trajectories and Clinical Outcomes Using PADRIS-PRESTO Population-Based Administrative Data

L’ús d’antipsicòtics en l’esquizofrènia a Catalunya: Un estudi d'efectivitat comparada en el món real de les trajectòries d'adherència i els resultats clínics utilitzant dades administratives de base poblacional de PADRIS-PRESTO

Barcelona, 10 de Juny de 2026

Author: Laia Raja Gubau

Director/s: Lina Leguízamo and Sergi Mas

Tutors: Lina Leguízamo and Sergi Mas

## Abstract

Medication non-adherence in schizophrenia remains one of the most significant barriers to effective treatment, contributing to relapse, psychiatric hospitalisation, and deteriorating quality of life. Long-acting injectable antipsychotics have been proposed as a strategy to improve adherence by eliminating the need for daily self-administration, yet real-world evidence on the longitudinal trajectory of adherence following a switch from oral formulations remains limited.

This study uses administrative dispensing data from the PADRIS-PRESTO cohort in Catalonia to characterise monthly adherence trajectories in patients with schizophrenia who transitioned from oral antipsychotics to long-acting injectables between 2015 and 2017. Time series clustering was applied to a final cohort of 1,025 patients, identifying four clinically distinct adherence trajectory groups: LAI Improvers, Chronically Adherent patients, Early Dropouts, and Moderate Improvers. Analysis of psychiatric healthcare contacts in the 12 months surrounding the switch revealed that LAI Improvers experienced the largest reductions in emergency department visits, psychiatric hospitalizations, and outpatient contacts.

These findings suggest that the clinical benefit of switching to long-acting injectables is strongly concentrated in patients with previously poor oral adherence, and that longitudinal trajectory analysis offers a more informative framework for evaluating treatment switches than conventional binary adherence metrics.


**Keywords**: Long-acting injectable antipsychotics (LAI), longitudinal trajectory of adherence, PADRIS-PESTO cohort, time series clustering, binary adherence metrics.

# Resum

L'incompliment terapèutic en l'esquizofrènia continua sent un dels principals obstacles per a un tractament efectiu, contribuint al risc de recaiguda, l'hospitalització psiquiàtrica i el deteriorament de la qualitat de vida dels pacients. Els antipsicòtics injectables de llarga durada s'han proposat com una estratègia per millorar l'adherència en eliminar la necessitat d'autoadministració diària, tot i que l'evidència en condicions de pràctica clínica real sobre l'evolució longitudinal de l'adherència després del canvi des de formulacions orals continua sent limitada.

Aquest estudi utilitza dades de dispensació farmacèutica de la cohort PADRIS-PRESTO de Catalunya per caracteritzar les trajectòries mensuals d'adherència en pacients amb diagnòstic d'esquizofrènia que van transitar d'antipsicòtics orals a injectables de llarga durada entre 2015 i 2017. Es va aplicar un algoritme de clustering de sèries temporals a una cohort final de 1.025 pacients, identificant quatre grups clínicament diferenciats: Milloradors amb LAI, Adherents crònics, Abandonadors primerencs i Milloradors moderats. L'anàlisi dels contactes psiquiàtrics en els 12 mesos al voltant del canvi va revelar que els milloradors amb LAI van experimentar les reduccions més grans en visites a urgències, hospitalitzacions psiquiàtriques i visites ambulatòries.

Aquests resultats suggereixen que el benefici clínic del canvi a injectables de llarga durada es concentra principalment en pacients amb adherència oral prèviament deficient, i que l'anàlisi de trajectòries longitudinals ofereix un marc més informatiu per avaluar els canvis de tractament que les mètriques binàries d'adherència convencionals.

**Paraules clau**: Antipsicòtics injectables de llarga durada (LAI), evolució longitudinal de l'adherència, la cohort PADRIS-PRESTO, clustering de sèries temporals, mètriques binàries d'adherència.

## Acknowledgements

I would like to thank Lina Leguízamo for guiding and helping me throughout this project, and Sergi Mas for listening to my idea and allowing me to pursue the project that I wanted to do. Ever since I took his pharmacology class, his teachings have always stayed with me.

I also want to specially thank Petter and Aaron for their generous help and for guiding me through the practical part of this project.

Finally, I want to thank my family, who has always been there for me whenever I struggled with this project and throughout these four years of university.

# Glossary

**AEMPS** Spanish Agency for Medicines and Medical Devices (Agencia Española de Medicamentos y Productos Sanitarios)

**ANOVA** Analysis of Variance

**AQuAS** Agency for Health Quality and Assessment of Catalonia (Agència de Qualitat i Avaluació Sanitàries de Catalunya)

**ATC** Anatomical Therapeutic Chemical (classification system)

**AIDS** Acquired Immunodeficiency Syndrome

**BMI** Body-mass index

**COPD** Chronic Obstructive Pulmonary Disease

**CIMA** Online Medicines Information Centre of the AEMPS

**DDD** Defined Daily Dose

**EMA** European Medicines Agency

**ICD-10** International Classification of Diseases, 10th revision

**LAI** Long-Acting injectable

**MPR** Medication Possession Ratio

**PADRIS** Health Research and Innovation Data Analytics Programme (Programa d'Analítica de Dades per a la Recerca i la Innovació en Salut)

**PDC** Proportion of Days Covered

**RWD** Real-World Data

**SISCAT** Comprehensive Public Use Health System of Catalonia **(**Sistema Sanitari Integral d'Utilització Pública de Catalunya)

**SMA** Outpatient mental health services (Salut mental Ambulatòria)

**SMH** Mental health unit (Salut mental Hospitalària)

**SmPC** Summary of Product Characteristics

**UMAP** Uniform Manifold Approximation and Projection

**URG** Emergency care (Urgències)

**WHO** World Health Organization

# List of Figures

Figure 1. Risk factors, neurobiology, and course of schizophrenia across the lifespan. ..... 5
Figure 2. Stepwise cohort selection from the PADRIS mental health database ..... 16
Figure 3. Silhouette score as a function of the number of clusters k, computed on the 2D UMAP mbedding ..... 25
Figure 4. 2D UMAP projection of the 24-month PDC trajectories ..... 25
Figure 5. Individual PDC trajectories for a random sample of 100 patients ..... 26
Figure 6. Heatmap of monthly PDC values for a random sample of 200 patients ..... 27
Figure 7. Violin plots of the monthly PDC distribution. ..... 27
Figure 8. Mean PDC trajectory with 95% confidence interval for Cluster 0 ..... 29
Figure 9. Mean PDC trajectory with 95% confidence interval and monthly psychiatric contact counts of Cluster 0. ..... 30
Figure 10. . Mean PDC trajectory with 95% confidence interval for Cluster 1 ..... 30
Figure 11. Mean PDC trajectory with 95% confidence interval and monthly psychiatric contact for Cluster 1. ..... 31
Figure 12. Mean PDC trajectory with 95% confidence interval for Cluster 2 ..... 31
Figure 13. Figure 12. Mean PDC trajectory with 95% confidence interval and monthly psychiatric contact total counts for Cluster 2 by service type ..... 32
Figure 14. Mean PDC trajectory with 95% confidence interval for Cluster 3 ..... 32
Figure 15. Mean PDC trajectory with 95% confidence interval and monthly total psychiatric contact counts for Cluster 3 by service type ..... 33
Figure 16. Summary comparison of average psychiatric contacts per patient across the four clusters , split by pre- and post-switch period ..... 34
Figure 17. Percentage of medication changes followed by a psychiatric contact ..... 35
Figure 18. Mean pre-switch PDC trajectory by main oral antipsychotic dispensed in the 12 months preceding the switch date and mean post-switch PDC trajectory by LAI formulation ..... 36
Figure 19. Gain-based feature importance from the Random Forest classifier ..... 38
Figure 20. SHAP beeswarm plots for each of the four trajectory clusters ..... 39
Figure 21. Predicted probability of LAI improvement and mean pre-switch PDC trajectory ..... 40
Figure 22. Predicted probability of LAI improvement and mean pre-switch PDC trajectory ..... 40
Figure 23. WBS diagram for the project ..... 42
Figure 24. Clear visualization of the overall plan ..... 43
Figure 25. PERT diagram ..... 44

# List of Tables

Table 1. . F20 schizophrenia subtypes identified in the dataset....4
Table 2. Classification of antipsychotics for Schizophrenia treatment.. ....7
Table 3. Structure and descriptive statistics of the synthetic dataset generated to simulate the PADRIS-PRESTO data.....20
Table 4. LAI antipsychotic products included in the study ....22
Table 5. Silhouette scores obtained for TimeSeriesKMeans solutions ....24
Table 6. Sample of the first ten patients in the monthly PDC....26
Table 7. Sociodemographic and clinical characteristics of the four adherence trajectory clusters identified by the k=4 TimeSeriesKMeans....28
Table 8. Test set performance metrics for the three tree-structured candidate models ....37
Table 9. Activities with their duration and precedents in hours. ....43
Table 10. SWOT analysis of the project ....46
Table 11. Breakdown of costs and resources for the project....47

# Contents

1. Introduction ..... 1
1.1. Background and Rationale ..... 1
1.2. Study Objectives ..... 1
1.3. Data Source: The PADRIS-Presto Cohort ..... 2
1.4. Scope and Limitations ..... 2
2. Background ..... 3
2.1. Schizophrenia: Epidemiology and Clinical Features ..... 3
2.2. Antipsychotic Treatment and Non-Adherence ..... 6
2.3. Long-Acting Injectable Antipsychotics ..... 7
2.4. Measuring Adherence from Electronic Health Records ..... 8
3. Market Analysis ..... 9
3.1. Current Landscape: Real-World Adhrence Research ..... 9
3.2. What This Project Contributes ..... 9
3.3. Prospects ..... 9
4. Conceptual Engineering ..... 11
4.1. Synthetic data generation ..... 11
4.1.1. Administrative and Legal Context ..... 11
4.1.2. Development of Synthetic Datasets ..... 11
4.1.3. Role and Limitations of the Synthetic Data ..... 12
4.2. Study of Solutions: Adherence Measurement ..... 12
4.2.1. Medication Possesion Ratio ..... 12
4.2.2. Binary Proportion of Days Covered ..... 12
4.2.3. Longitudinal PDC Trajectory Analysis ..... 13
4.3. Antipsychotic selection and formulation classification ..... 14
4.3.1. Selection of Antipsychotic Agents ..... 14
4.4. PDC Calculation design ..... 15
4.5. Cohort definition for trajectory analysis ..... 16
4.6. Time series clustering ..... 17
4.7. Predictive modelling: Identifying candidate patients for LAI switching ..... 17
5. Detail Engineering ..... 19
5.1. Synthetic Data Generation ..... 19
5.2. Analytical Pipeline ..... 21

5.2.1. Data Sources and Tables ..... 21
5.2.3. Drug Formulation Classification ..... 21
5.2.4. PDC Calculation ..... 22
5.2.5. Cohort Definition and Patient Selection ..... 23
5.2.7. Adherence Trajectory Clusters and Healthcare Contact Patterns ..... 26
5.2.9. Toxicity Episode Detection ..... 34
5.2.10. PDC Trajectory by Drug type ..... 35
5.3. Predictive Modelling: Identifying candidate patients for LAI switching ..... 36
5.3.1. Feature Matrix Construction ..... 36
5.3.2. Hyperparameter Optimization and Model Evaluation ..... 37
5.3.3. Feature Importance Analysis ..... 38
5.3.4. SHAP Analysis by Cluster ..... 38
5.3.5. Prospective Application to Oral-Only Patients ..... 39
6. Execution Schedule ..... 42
6.1 Work Breakdown Structure (WBS) ..... 42
6.2 Execution Chronogram (GANTT) ..... 42
6.3 Critical Path Method Diagram (PERT) ..... 44
7. Technical Viability ..... 45
8. Economic Viability ..... 47
9. Regulations and Legal Aspects ..... 48
10. Conclusions and Future Lines ..... 49
11. Bibliography ..... 51

# 1. Introduction

## 1.1. Background and Rationale

Schizophrenia is a severe and chronic psychiatric disorder that affects approximately 1% of the global population [1] and represents one of the leading causes of disability worldwide. Its complex symptomatology, including positive symptoms such as hallucinations and delusions, negative symptoms such as diminished emotional expression and loss of motivation, and significant cognitive deficits, requires long-term pharmacological treatment and ongoing clinical monitoring.

Therapeutic adherence is directly associated with clinical outcomes in schizophrenia. Poor adherence, which remains one of the most critical challenges in schizophrenia care, has been consistently associated with an increased risk of relapse [2], psychiatric hospitalization, and deterioration in social and occupational functioning. In contrast, sustained adherence to antipsychotic treatment is associated with reduced symptom severity, lower hospitalization rates, and improved quality of life.

Two principal modalities of antipsychotic administration are relevant to adherence research: oral formulations, including tablets, capsules, and oral solutions, and injectable formulations, which are further divided into immediate-effect preparations used primarily in acute crisis management and long-acting injectable formulations (LAIs), administered at intervals ranging from two weeks to three months depending on the specific compound.

LAIs were developed to address the limitations of oral treatment, particularly poor adherence, as their administration in a clinical setting allows for a clearer identification of non-adherence [3]. In addition, this mode of administration reduces the likelihood of unrecognized or covert non-adherence compared to oral treatment. However, their comparative effectiveness in real-world settings remains under investigation, and evidence from population-based studies is still limited [4].

## 1.2. Study Objectives

This thesis aims to develop and apply a data analysis pipeline to determine and compare the therapeutic adherence of patients diagnosed with schizophrenia receiving long-acting injectable (LAI) versus oral antipsychotic treatments in Catalonia, using population-based electronic health records from the Catalan Health System.

1. To identify and characterize a cohort of patients with a confirmed diagnosis of schizophrenia (ICD-10: F20) within the PADRIS-PRESTO database between 2010 and 2019.
2. To evaluate longitudinal adherence using the Proportion of Days Covered (PDC) metric for both LAI and oral antipsychotic users by computing monthly adherence values and exploring adherence patterns through time-series clustering techniques.
3. To compare adherence to relevant sociodemographic and clinical covariates, including age, sex, socioeconomic level, healthcare region, and medical comorbidities.
4. To explore the association between adherence patterns and healthcare resource utilization outcomes such as psychiatric emergency visits, and primary and ambulatory care contacts.
5. To develop an exploratory proof-of-concept predictive algorithm to anticipate whether switching a patient to an LAI is likely to produce a meaningful adherence improvement.

## 1.3. Data Source: The PADRIS-Presto Cohort

The primary data source for this study is the PADRIS-PRESTO cohort, a comprehensive population-based database covering over 1.4 million individuals registered in the Catalan Health System between 2015 and 2019, with retrospective follow-up extending to 2010. PADRIS, which stands for Programme for the Analytics of Health Data, is a data analytics programme developed by the Agency for Health Quality and Assessment of Catalonia (AQuAS) [5], which enables the linkage and re-use of health data generated by the SISCAT (Sistema Sanitari d'Utilització Pública de Catalunya) [6] in accordance with applicable legal and ethical frameworks.

Beyond mental health, the PADRIS programme supports a broad portfolio of population-based research projects, including studies on frailty in older adults, and operates under the legal, ethical, and transparency frameworks governing the re-use of data generated by the SISCAT.

The PRESTO (Primary Care Digital Support Tool) extension of PADRIS integrates longitudinal individual-level records across multiple domains of the healthcare system, including primary care, specialized mental health outpatient services, hospital admissions, emergency departments, and community pharmacy dispensations. The cohort includes 473,812 individuals with at least one contact with specialized mental health services (study population) and 947,698 individuals without a recorded psychiatric diagnosis in specialized services (control population).

The PADRIS-PRESTO cohort was constructed using data from the PADRIS program by Michele De Prisco and Vincenzo Oliva, researchers from the Bipolar and Depressive Disorders group at IDIBAPS at Hospital Clínic of Barcelona, in collaboration with AQuAS, under the supervision of Eduard Vieta, Diego Hidalgo-Mazzei, and Gerard Anmella. It represents one of the first population-scale analyses of diagnostic patterns, treatment trajectories, and healthcare service utilization in mental health in Catalonia [6].

Key findings from the cohort's foundational publication underscore the clinical relevance of this data source: individuals with mental health diagnoses have been reported to utilize healthcare services at higher rates compared to those without psychiatric diagnoses, with notably higher emergency department attendance. Relevant clinical indicators include elevated rates of smoking, greater economic vulnerability, and an increased prevalence of physical comorbidities such as musculoskeletal disorders, hypertension, and diabetes. Over the ten-year follow-up period, 67 million prescriptions were recorded [6], with antidepressants and anxiolytics accounting for the largest share, highlighting the importance of reviewing prescription patterns to ensure alignment with clinical indicators.

## 1.4. Scope and Limitations

This study operated within a well-defined scope. The analytical cohort is restricted to individuals with a recorded diagnosis of Schizophrenia (ICD-10 F20.x) who received at least one antipsychotic prescription (ATC: N05A) within the PADRIS system between 2010 and 2019, allowing for a minimum two-year follow-up window. The analysis draws on anonymous pharmacy dispensing records, primary care visit registries, psychiatric outpatient service records, hospitalization data, and emergency department records.

# 2. Background

## 2.1. Schizophrenia: Epidemiology and Clinical Features

Schizophrenia is a severe, debilitating, and chronic psychiatric disorder that affects approximately 1% of the population worldwide. However, the timing of onset differs by sex: men tend to develop the disorder in their late teens or early twenties, while women typically present later, in their late twenties or early thirties [7], with a second smaller peak around menopause (44–49 years). The lower incidence in premenopausal women is explained by the protective effect of estrogen on dopamine receptors [8].

It is characterized by positive and negative symptoms that can influence patients' thoughts, perceptions, speech, affect, and behaviors. Positive symptoms include hallucinations and delusions that are commonly paranoid, and negative symptoms include loss of the ability to experience pleasure (anhedonia), loss of will, social withdrawal, and overall flattened affect [9]. Cognitive deficits, including impairments in memory, attention, and executive function, are also recognized as a core dimension of the disorder, distinct from both positive and negative symptom domains [9]. Together, these symptoms profoundly affect not only the individual but also their immediate social and family environment, highlighting the need for early and sustained clinical intervention [7].

In addition to positive and negative symptoms, schizophrenia is associated with significant social and occupational dysfunction. One or more key areas of daily functioning, such as work, relationships, or self-care, are noticeably below the level the individual has previously achieved. In the case of an early development of this illness, this may instead manifest as a failure to develop age-appropriate functioning in social, academic, or occupational domains.

Schizophrenia is also characterized by disorganized thought, which manifests in both speech and behavior. Disorganized speech can range from loose associations and rapid topic-switching to severely incoherent output known as schizophasia, in which words are used in a confused, repetitive manner with no apparent logical connection. Disorganized behavior, on the other hand, can interfere with basic daily activities such as meal preparation or personal hygiene, and may also present as inappropriate silliness or episodes of unpredictable agitation. It is worth noting that no single sign or symptom is pathognomonic of the disease [7]. A definitive diagnosis requires that symptoms be present for a significant portion of at least one month, with some signs persisting for a minimum of six months.

The disorder is further classified into subtypes according to the ICD-10 coding system. The following subtypes were identified in the present dataset:

| ICD-10 Code | Type of Schizophrenia | Description |
|---|---|---|
| F20.0 | Paranoid | Dominated by delusions and auditory hallucinations; cognitive function and affect relatively preserved |
| F20.1 | Disorganized | Disorganized speech and behavior, with flat or inappropriate affect |
| F20.2 | Catatonic | Motor disturbances ranging from immobility and stupor to excessive purposeless activity and negativism |
| F20.3 | Undifferentiated | Meets general criteria but does not fit the previous subtypes |

| | | |
|---|---|---|
| F20.5 | Residual | Ongoing negative symptoms and attenuated positive symptoms in the absence of significant active psychosis |
| F20.81 | Schizophreniform disorder | Shorter duration of symptoms |
| F20.89 | Other types | Do not fit any of the above subtypes |
| F20.9 | Unspecified | Insufficient information to classify into a specific subtype |

*Table 1. . F20 schizophrenia subtypes identified in the dataset. Source: own elaboration [7].*

The most significant known risk factor is a family history of schizophrenia [10]. Other factors such as season and location of birth, socioeconomic status, and maternal infections during pregnancy have been proposed, though the evidence supporting these associations remains inconclusive.

It is understood to be a polygenic disorder, meaning that multiple genes interact with environmental and developmental factors to influence an individual's likelihood of developing the illness [11]. What remains unclear is whether patients present so differently from one another because the disease affects different parts of the brain, has different underlying causes, or because what we call "schizophrenia" may actually be an umbrella term for several related but distinct conditions that look similar on the surface. This diagnostic ambiguity is compounded by the fact that schizophrenia is currently diagnosed based on clinical symptoms alone, in the absence of validated biological markers, which limits the precision with which patients can be classified and treated. The heterogeneity of the disorder therefore represents a fundamental challenge for clinical practice. It underscores the need for patient stratification based on biological and phenotypic markers; an opportunity that precision medicine approaches, including machine learning applied to large-scale clinical data, are beginning to address [12].

The Finnish Adoptive Family Study of Schizophrenia confirmed that genetics plays a central role in the development of the disorder. Interestingly, the study also found that individuals who carry a genetic risk for schizophrenia are especially sensitive to their family environment; growing up in a home with little criticism and open, clear communication seems to reduce the chances of the illness actually developing [13]. More recent evidence corroborates this gene-environment interplay, showing that polygenic risk interacts with environmental exposures across the psychosis spectrum and that these interactions can be detected even in first-episode cohorts [14, 15, 16].

At a neurobiological level, the dopaminergic system plays a key role. Substances that induce psychotic states similar to the positive symptoms of schizophrenia tend to increase dopaminergic neurotransmission in the brain, while virtually all antipsychotic medications work by reducing it [1].

Beyond triggering transient psychotic states, chronic substance use, particularly cannabis and stimulants, has also been associated with an increased long-term risk of developing schizophrenia in genetically vulnerable individuals.

The diagnosis of this disease also requires ruling out other potential explanations for the symptoms. If psychotic features can be attributed to substance use, such as stimulant or hallucinogen consumption, medication side effects, or an underlying medical condition, a diagnosis of schizophrenia cannot be made, even when the clinical presentation appears identical [10]. This exclusion criterion is particularly relevant given that, as mentioned above, certain substances that increase dopamine activity in the brain can trigger psychotic states that closely resemble the positive symptoms of schizophrenia, highlighting the importance of a thorough clinical assessment before reaching a definitive diagnosis.

The clinical course of schizophrenia is generally understood through three main management phases: the acute phase, characterized by the emergence or worsening of psychotic symptoms; the stabilization phase, during which symptoms begin to subside and the patient starts to recover from the acute episode; and the maintenance phase, focused on preventing relapse and sustaining the functional gains achieved [Fig. 1].

In many patients, these phases are preceded by a prodromal period, during which the illness does not appear suddenly but rather builds up gradually over time. During this period, subtle changes in cognition, emotion, perception, and sleep begin to emerge, often going unrecognized by both the patient and those around them. This prodromal stage can last anywhere from a few weeks to several years before full psychosis develops, which means that by the time a formal diagnosis is made, the disorder may have already been silently progressing for a considerable amount of time [17]. Critically, the clinical course is strongly conditioned by medication adherence. Patients who experience multiple relapses over time but retain a biological response to antipsychotics are frequently those with poor adherence to oral treatment, and it is precisely this subgroup that stands to benefit most from long-acting injectable formulations [18].

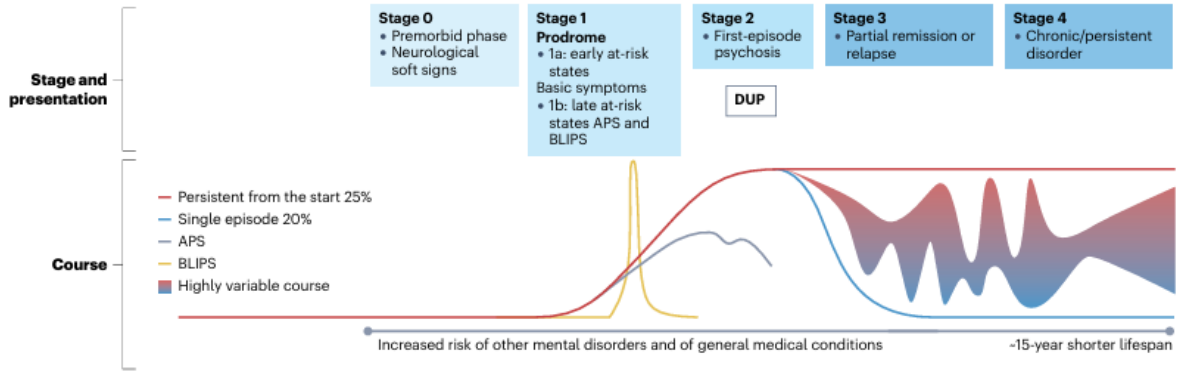



*Figure 1. Risk factors, neurobiology, and course of schizophrenia across the lifespan. Adapted from Leucht et al. (2025) [18].*

The economic burden of schizophrenia in Europe is considerable. An estimated four million people across Europe live with schizophrenia, with a point of prevalence of 0.25% to 0.56% and an incidence of 15.2 per 100,000 people [19]. Across European countries, the annual per-patient cost varies substantially depending on the healthcare system and methodology used, ranging from 6,028€ in Spain to 23,241€ in the United Kingdom [19]. In Spain specifically, the total cost of schizophrenia has been estimated at nearly €2 billion, including direct medical costs (hospitalizations, outpatient consultations, and medication) and informal caregiving costs. This places schizophrenia at approximately 2.7% of total public healthcare expenditure in the country [20] and people with mental disorders use healthcare services at nearly twice the rate of the general population [21].

Across Europe more broadly, indirect costs driven by unemployment, lost productivity, and unpaid family caregiving account for between 50% and 90% of the total burden yet remain systematically underreported in healthcare planning [22]. Schizophrenia is not only a clinical challenge but a major socioeconomic one, and any intervention capable of reducing relapse rates and maintaining treatment continuity carries significant economic implications beyond the clinical setting.

## 2.2. Antipsychotic Treatment and Non-Adherence

Traditionally, schizophrenia has always been managed through daily oral antipsychotics. However, this treatment presents a fundamental challenge: its effectiveness is highly dependent on patient adherence to a consistent administration routine.

Beyond management of psychological symptoms, patients with schizophrenia frequently experience poorer oral health outcomes compared to the general population, with antipsychotic medications being a potential contributing factor [23]. The most commonly reported oral health complications include disturbances in salivary flow, tooth loss, periodontal disease and orofacial dyskinesia, which is a neurological disorder characterized by involuntary, repetitive, and abnormal movements of the facial and oral muscles [24]. The relationship between oral health and medication adherence is particularly significant. When patients experience oral pain or persistent discomfort, they may begin to perceive their treatment as more harmful than beneficial, which can ultimately lead to intentional non-adherence.

By the time schizophrenia is formally diagnosed, a significant amount of damage has already been done. The prodromal phase, which can stretch over several years before the first acute episode, is precisely the period where the illness takes its greatest social toll, disrupting education, relationships, and daily functioning before anyone has even started treatment. This makes a strong case for intervening early and decisively once a diagnosis is confirmed [8].

Over a follow-up of 11.2 years, the number of psychotic relapse episodes ranged from 0 to 29, with a mean of 3. Positive symptoms produced the highest number of relapses and the shortest episodes (mean 2 months), while negative and depressive symptoms showed longer exacerbations (3–5 years) [8].

Long-term treatment adherence remains one of the greatest challenges of all chronic medical conditions. It is a particular challenge in schizophrenia, given its association with social isolation, stigma, and comorbid substance misuse. Nonadherence affects more than one-third of patients with schizophrenia annually [2]. Irregular intake of oral medication frequently leads to symptom relapse, increased hospitalization rates, self-harm, and poor long-term positive effects, all of which contribute to a reduced quality of life. Further complicating this issue is the absence of an accepted gold standard for the measurement of adherence, as every available method has its limitations and advantages.

Antipsychotic medications are broadly divided into two generations. First-generation antipsychotics (FGAs) were the first pharmacological treatment developed for schizophrenia and work primarily by blocking dopamine D2 receptors. While effective for positive symptoms, their use is associated with a high incidence of extrapyramidal side effects (EPS) and tardive dyskinesia, which frequently contribute to treatment discontinuation. Second-generation antipsychotics (SGAs), or atypical antipsychotics, emerged as an alternative with a broader receptor profile, also targeting serotonin receptors, resulting in a lower risk of EPS and akathisia. However, SGAs are associated with greater weight gain and metabolic side effects [25].2

| Generation | Type | Examples | Primary Receptor Target | Main Side Effects | Forms of Administration |
|---|---|---|---|---|---|
| First Generation (FGA) | Typical / Conventional | Haloperidol, Chlorpromazine, Sulpiride, Zuclopenthixol | D2 dopamine antagonist | EPS, tardive dyskinesia, akathisia | Oral, short-acting injectable, LAI |
| Second Generation (SGA) | Atypical | Olanzapine, Risperidone, Quetiapine, Aripiprazole, Clozapine, Amisulpride, Paliperidone | D2+ serotonin (5-HT2A) antagonist | Weight gain, metabolic effects, sedation | Oral, short-acting injectable, LAI |

*Table 2. Classification of antipsychotics for Schizophrenia treatment. Source: own elaboration **[25]**.*

A meta-analysis of thirteen randomized controlled trials in first-episode schizophrenia found that pooled SGAs were associated with significantly lower treatment discontinuation rates compared to FGAs, reducing all-cause discontinuation by 26%, discontinuation due to inefficacy by 40%, and discontinuation due to intolerability by 54%. SGAs also outperformed FGAs on negative symptoms and global cognitive function [25]. These findings are particularly relevant in the context of adherence, as treatment discontinuation is one of the strongest proxies for non-adherence in real-world clinical settings.

Notably, both generations include long-acting injectable formulations [Table 2], meaning that it is not limited to a single pharmacological class but reflects a difference in administration modality that applies across the full antipsychotic landscape.

Long-term antipsychotic treatment is necessary for all patients with schizophrenia, even though dosage management during the stable phase presents considerable challenges. High-maintenance doses, particularly above 600 mg CPZ equivalents for first-generation antipsychotics, have not been shown to offer superior relapse prevention over standard doses. These complexities are compounded by the fact that oral antipsychotic regimens are entirely dependent on patient self-administration, making adequate dosing difficult to guarantee in clinical settings. Ensuring treatment continuity with oral medication remains one of the central limitations of this treatment modality [26].

## 2.3. Long-Acting Injectable Antipsychotics

Long-acting injectable antipsychotics (LAIs) have emerged as a strategic alternative to daily oral regimens. By extending the dosing interval, ranging from every two weeks to once every three months, LAIs may improve adherence and reduce relapse compared to oral antipsychotics, while also ensuring stable plasma drug concentrations and significantly reducing the burden of daily pill-taking.

LAIs have been associated with several potential advantages over oral antipsychotics (OAs), including enhanced medication adherence, reduced risk of relapse and hospitalization, and the capacity to distinguish between genuine treatment resistance and “pseudo”-resistance, which refers to the inability of the drug to be efficient because of the patient not following its posology. Long-acting medications have been associated in some studies with lower mortality compared to OAs [3].

Furthermore, as LAIs are administered by a healthcare professional, missed doses can be more readily identified compared to oral treatment, allowing clinicians and caregivers to intervene without delay and support patients in maintaining their treatment schedule. This supervised administration also provides clinicians with reliable data to ensure adherence, evaluate efficacy, and determine whether the medication is producing the intended therapeutic response [27]. Current guidelines now recommend that healthcare professionals discuss the treatment with LAIs first with patients and use them in the early stages of the course of the illness, rather than as a last resort.

This supervised administration also facilitates the routine clinical monitoring that guidelines recommend during the stable phase, including assessment of extrapyramidal symptoms (EPS), weight gain, and cardiovascular and metabolic side effects. Special attention should be given to obesity-related complications, with appropriate interventions considered where necessary [26].

Despite all these benefits, LAIs still remain underutilized across many healthcare settings. This is due to negative attitudes, misconceptions, and insufficient knowledge among clinicians and patients [3]. Another practical barrier is the acquisition of costs of these medications and the lack of adequate service infrastructure for their administration and monitoring.

Unlike oral formulations, where the DDD is derived from daily intake, LAIs release the active compound gradually following intramuscular injection, achieving stable plasma concentrations over an extended period. The duration of coverage for each LAI formulation is determined by its pharmacokinetic profile, specifically the absorption rate from the injection site and the half-life of the active compound and is standardized in the Summary of Product Characteristics (SmPC) for each drug.

## 2.4. Measuring Adherence from Electronic Health Records

Electronic health records and administrative claims databases have become the primary data source for population-level adherence research, offering large sample sizes, long follow-up periods, and real-world clinical data that controlled trials cannot provide. Within this framework, two aggregate metrics have been established as the standard for measuring medication adherence: the Medication Possession Ratio (MPR) and the Proportion of Days Covered (PDC).

The MPR is calculated as the total number of days' supply of medication dispensed divided by the number of days in the observation period. The PDC refines this approach by capping overlapping supplies at the date of the next dispensation, avoiding overestimation of adherence when patients collect medication early. Both metrics produce a value between 0 and 1, with a threshold of 0.80, or 80%, conventionally used to classify a patient as adherent [4]. Importantly, both metrics estimate medication availability rather than actual medication intake.

A patient with an average PDC of 0.75 may have been fully adherent for the first six months and completely non-adherent for the last six, or consistently moderately adherent throughout, two clinically very different situations that the aggregate metric cannot distinguish. This limitation has driven the adoption of Group-Based Trajectory Modeling (GBTM) as a more informative analytical approach. GBTM is a statistical method that identifies groups of patients with similar longitudinal adherence patterns over time, rather than summarizing behavior as a single average [28].

# 3. Market Analysis

## 3.1. Current Landscape: Real-World Adherence Research

The evaluation of antipsychotic adherence using real-world data has become an increasingly active field within pharmacoepidemiology, driven by the recognition that findings from randomized controlled trials, conducted under highly controlled conditions with selected patient populations, do not always translate to routine clinical practice. Real-world studies using administrative claims and dispensing databases have consistently demonstrated that adherence to oral antipsychotics in schizophrenia is poor, with only around one-third of patients maintaining consistent medication use and discontinuation rates approaching 80% within the first year of treatment [4].

Comparative analyses of LAI versus oral formulations in real-world cohorts have produced consistent findings in favor of LAIs, particularly in non-adherent patients. A nationwide mirror-image study conducted in France using the French National Health Data System, which is a population-based database structurally comparable to the PADRIS system used in the present study, enrolled 12,373 patients with schizophrenia initiating LAI treatment and found that 63.7% were non-adherent to their previous oral antipsychotic regimen at the time of switching. In this non-adherent group, LAI initiation was associated with a 20 to 26% reduction in psychiatric hospitalizations and a 29% decrease in emergency department admissions in the year following initiation, compared to the year prior. Crucially, this benefit was largely absent in patients who had already been adherent to oral treatment, except for aripiprazole LAI, which showed a modest reduction in hospitalizations even in compliant patients [4]. These findings directly motivate the central objective of the present study: to identify, from population-level dispensing data, which patients within the Catalan health system are most likely to benefit from a switch to LAI formulations.

Population-based research using Southern European healthcare data remains scarce, and no published study has applied trajectory-based methods to characterize how adherence evolves within the Catalan public health system. This project aims to address that gap.

## 3.2. What This Project Contributes

This project contributes to the pharmacoepidemiology of antipsychotic treatment in Catalonia by leveraging real-world administrative dispensing data from PADRIS to evaluate LAI adherence patterns at a population level. Unlike controlled clinical trials, which operate under ideal conditions, this study captures how LAI antipsychotics perform in routine clinical practice across the Catalan healthcare system (CatSalut). Developed with the Pharmacology unit in the Hospital Clínic de Barcelona in collaboration with the AI laboratory at IDIBAPS, the work translates large-scale, heterogeneous population data into structured, interpretable insights, providing regional evidence on LAI effectiveness that can inform prescribing practices and healthcare resource planning.

## 3.3. Prospects

The pipeline developed in this project is designed to be modular and reproducible, meaning it can be adapted beyond its current application with relatively limited additional development. Several directions present clear clinical and institutional value.

In the short term, the methodology could be extended to other chronic psychiatric conditions managed within PADRIS, such as bipolar disorder or treatment-resistant depression, where long-term pharmacological adherence is equally critical and similarly difficult to monitor.

In the medium term, the adherence trajectory outputs could be linked to clinical outcome data, including hospitalization rates, emergency department visits, and mortality, to formally test whether the trajectory groups identified in this study predict meaningful differences in patient outcomes, which would strengthen the evidence base for LAI prioritization policies.

In the longer term, the dispensing pattern analysis could inform hospital pharmacy planning by providing real-world data on LAI consumption trends across healthcare regions, supporting more efficient stock management and resource distribution across the Catalan public health network.

# 4. Conceptual Engineering

## 4.1. Synthetic data generation

### 4.1.1. Administrative and Legal Context

The acquisition of real patient data from PADRIS (Programa d'Analítica de Dades per a la Recerca i la Innovació en Salut), accessed through the Psychiatry Unit at Hospital Clínic de Barcelona, was subject to rigorous anonymization protocols and a formal legal approval process. Given the sensitive nature of psychiatric patient records, the administrative procedure for data release extended over several months. Rather than halting development during this waiting period, a synthetic dataset was generated to serve two parallel purposes: validating the analytical pipeline in a controlled environment and enabling early-stage algorithm development before the real data became available.

### 4.1.2. Development of Synthetic Datasets

The synthetic dataset was constructed through a comprehensive review of published studies using PADRIS data. These sources provided the most relevant variables and the statistical benchmarks necessary to simulate a realistic population, including variable coverage rates, central tendency measures such as median and mean values, standard deviations, and categorical frequency distributions.

Using these parameters, synthetic datasets were created across five categories to simulate the structure of a PADRIS extract. The data generation process was implemented in Python using the pandas and numpy libraries for data structuring and statistical simulation, with all generated files stored in CSV format. The generation process was executed in two steps: first, a pilot dataset of 1,000 synthetic individuals was created to validate the internal consistency of the generated variables and subsequently scaled up to 100,000 subjects using extrapolation techniques to better reflect the complexity of the full study population of 1,421,510 patients. Following a non-control study design, all generated patients were assigned a documented history within a mental health unit.

The dataset was organized into five complementary files, each simulating a distinct category of PADRIS data. The first covered individual sociodemographic and clinical profiles, including variables such as sex, age, socioeconomic level, nationality, healthcare region, and smoking status. The second added a clinical risk layer with comorbidity counts and binary indicators for specific chronic conditions, parameterized to match prevalence frequencies reported in the literature. The third simulated community pharmacy dispensing records, structured around the ATC drug classification system, capturing the type, quantity, and cost of medications dispensed per patient per month. The fourth reproduced healthcare contact records across different care settings, including primary care, outpatient mental health, and hospitalization episodes. The fifth contained psychiatric diagnostic information, assigning each patient one or more diagnostic categories with their corresponding ICD-10 codes, which was essential for testing the cohort selection logic before accessing the real data.

This structure was deliberately designed to mirror the organization of the real PADRIS tables so that the analytical pipeline could be developed and validated in a controlled environment before the actual data became available.

### 4.1.3. Role and Limitations of the Synthetic Data

While the synthetic dataset enabled early pipeline testing and the development of core preprocessing and adherence calculation routines, it presented an important limitation: because it was derived from population-level summary statistics rather than individual-level records, it lacked the clinical specificity of the real PADRIS data. In particular, the synthetic dispensing records could not faithfully reproduce the complex longitudinal patterns of antipsychotic switching, the heterogeneity of LAI injection intervals across products, or the irregularity of real-world dispensing behavior.

For this reason, once the PADRIS data was accessible, all subsequent analyses were conducted exclusively on the real dataset. The synthetic data is nevertheless reported here as a methodological step that allowed development work to proceed in parallel with the administrative approval process, reducing the overall project timeline.

## 4.2. Study of Solutions: Adherence Measurement

Measuring medication adherence from administrative dispensing data is not a straightforward task. Several methodological approaches exist in the literature, each with different assumptions, levels of granularity, and suitability for longitudinal analysis. Three main alternatives were considered for this study: the Medication Possession Ratio, the binary Proportion of Days Covered, and the longitudinal PDC trajectory approach.

### 4.2.1. Medication Possession Ratio

The Medication Possession Ratio (MPR) is one of the oldest and most widely used adherence metrics in pharmacoepidemiology. It is calculated as the total number of days of medication supply dispensed divided by the total number of days in the observation period. Its main advantage is computational simplicity; it requires only a count of dispensed units and a defined follow-up window.

However, MPR has a well-known limitation: it can exceed 1.0 when a patient stockpiles medication through early refills, producing values that are difficult to interpret clinically and that inflate apparent adherence. More critically for this study, MPR produces a single summary statistic per patient per observation period, which means it entirely collapses the temporal dimension of adherence. A patient who takes their medication consistently for six months and then stops completely would receive the same MPR as a patient who alternates between adherent and non-adherent months throughout the year. For a study whose central question is how adherence patterns evolve before and after a treatment switch, this loss of temporal information makes MPR fundamentally inadequate.

### 4.2.2. Binary Proportion of Days Covered

The Proportion of Days Covered (PDC) addresses MPR's ceiling problem by capping coverage at 1.0 per time unit and accounting for early refill overlap by carrying forward leftover days rather than double-counting them. It is now considered the preferred adherence metric by most pharmacoepidemiological guidelines.

In its most common form, PDC is computed over a fixed observation window and then converted to a binary variable at a threshold of 80%, classifying patients as either adherent or non-adherent. This binary approach has good clinical face validity, as the 80% threshold is widely used in the literature as a proxy

for therapeutic effectiveness, and it allows straightforward group comparisons and logistic regression analyses.

Large-scale real-world studies have adopted this approach precisely. Boyer et al. [4], in a nationwide mirror-image study of 12,373 patients with schizophrenia drawn from the French National Health Data System, defined compliance as exposure to an oral antipsychotic for at least 80% of the year preceding LAI initiation, and demonstrated that LAI initiation was, in this approach precisely, reducing psychiatric hospitalizations and emergency department admissions specifically in non-compliant patients. While this study represents an important contribution to the field and serves as a key methodological reference for the mirror-image design adopted here, it also illustrates precisely the limitation of the binary PDC approach: by collapsing the entire pre-switch year into a single label that is compliant or non-compliant, it cannot distinguish between a patient who was chronically poorly adherent and one who maintained good adherence for nine months before an acute deterioration, nor can it capture whether non-adherence was stable, worsening, or improving over time. Two patients with identical binary labels can have completely different longitudinal profiles, and these distinctions are clinically meaningful.

Additionally, when the research question involves comparing adherence before and after a specific index event, as is the case here, a single global PDC value provides no insight into the direction or magnitude of change around that event, nor into the rate at which adherence responds to the treatment switch.

#### 4.2.3. Longitudinal PDC Trajectory Analysis

The approach selected for this study computes PDC at monthly resolution for each patient across a 24-month window centered on the date of the oral-to-LAI switch, covering 12 months before and 12 months after. This produces a continuous adherence trajectory per patient rather than a single summary statistic, preserving the full temporal structure of medication-taking behavior. MacEwan et al. [29] applied a similar trajectory-based approach to longitudinal PDC data from over 29,000 patients with schizophrenia on oral atypical antipsychotics, identifying six clinically distinct adherence trajectory groups and demonstrating that classifying patients based on a single average PDC measure masks meaningful variations over time in how patients arrive at a given adherence level. This is the central methodological argument for the approach adopted in the present study.

The key advantage of this approach in the context of this study is that it allows the analysis to capture not just whether adherence improved after switching, but how it improved, for whom, and at what rate. Patients who were already adherent to oral medication before switching behave very differently from patients who were poorly adherent and then dramatically improved on LAI. In addition, both of these groups are clinically distinct from patients who showed a brief initial improvement before dropping out entirely. None of these patterns is recoverable from a binary or summary metric.

Furthermore, computing PDC separately for oral and LAI formulations within each month allows the transition period itself to be characterized, rather than treating the switch as a clean boundary. These implementation choices are not arbitrary but reflect a series of methodological decisions grounded in both clinical reasoning and the constraints of administrative dispensing data.

## 4.3. Antipsychotic selection and formulation classification

### 4.3.1. Selection of Antipsychotic Agents

The antipsychotic agents included in this study were defined in accordance with the research protocol established by the FRCB-IDIBAPS pharmacoepidemiology group at Hospital Clínic de Barcelona, within the framework of the PADRIS-PRESTO cohort study. The selection covers eleven compounds belonging to the ATC N05A class, chosen to represent the full spectrum of antipsychotic treatments in current clinical use in Catalonia: haloperidol (N05AD01), clozapine (N05AH02), olanzapine (N05AH03), quetiapine (N05AH04), clothiapine (N05AH06), sulpiride (N05AL01), amisulpride (N05AL05), risperidone (N05AX08), aripiprazole (N05AX12), paliperidone (N05AX13), and lurasidone (N05AE05). Together, these compounds account for the vast majority of antipsychotic dispensing events in the Catalan population with a schizophrenia diagnosis, spanning both first-generation agents and the full range of second-generation compounds across different receptor binding profiles.

The selection covers both first-generation agents, represented here by haloperidol, and the full range of second-generation compounds, reflecting the actual prescribing landscape in Catalonia during the study period. This scope ensures that the cohort is not artificially restricted to a particular pharmacological class, which would limit the generalizability of the adherence findings.

### 4.3.2. Formulation Classification

Once the set of eligible antipsychotic agents was defined, each dispensed product needed to be classified into one of three formulation categories: oral, long-acting injectable, and immediate-action injectable. This classification determines how days of coverage are calculated for each dispensing event and therefore has a direct impact on the PDC computation.

Oral formulations include all tablets, capsules, and oral solutions. These are the reference formulations in this study, representing standard community pharmacy dispensing where the patient self-administers the medication daily. For oral products, the number of days covered by a dispensing event is derived directly from the dispensed quantity expressed in Defined Daily Doses.

Long-acting injectable formulations are the intervention of interest. Unlike oral medications, LAIs are administered by a healthcare professional at fixed intervals ranging from two weeks to three months, depending on the product. Each LAI product was therefore assigned at a fixed injection interval based on the Summary of Product Characteristics, as approved by the European Medicines Agency and registered in the CIMA database.

Immediate-action injectable formulations represent a distinct category of short-acting intramuscular preparations used in acute psychiatric settings, and they do not contribute to ongoing medication coverage in the way that oral or LAI formulations do. They were therefore excluded from the PDC calculation while being retained in the dataset for other analytical purposes.

## 4.4. PDC Calculation design

### 4.4.1. Defining the Observation Window and Index Date

The index date for each patient is defined as the date of their first LAI dispensing event. Since PADRIS only records dispensing at year-month resolution rather than exact dates, the 15th of the dispensing month was adopted as a proxy for the actual dispensing date across all records. This mid-month imputation is a standard convention in pharmacoepidemiological studies using administrative data, as it minimizes the expected absolute error compared to any other fixed-day assumption.

The observation window spans 24 months centered on the index date: months -12 to -1 constitute the pre-switch period, during which the patient was on oral antipsychotics, and months +1 to +12 constitute the post-switch period, during which LAI treatment was initiated. The switch month itself is excluded from both windows, as it represents a transition period where oral and LAI coverage may overlap, and the PDC signal is ambiguous. Only patients with complete PDC data across the full 24-month window were included in the trajectory analysis cohort.

### 4.4.2. Days Covered: Oral Formulations

For oral formulations, the number of days covered by a dispensing event is taken directly from the Daily Defined Dose (DDD), which records all dispensed across all units in that event. This field effectively encodes the expected duration of treatment for that dispensing, making it the natural numerator for the PDC calculation.

A key design decision here concerns early refills. In real-world dispensing, patients often collect a new supply before their previous supply has run out, either because of appointment timing or as a precautionary buffer. The standard approach, adopted here, is to carry forward any leftover days from the previous dispensing event before starting the coverage clock for the new one.

### 4.4.3. Days Covered: LAI Formulations

For LAI formulations, the days-covered calculation follows a different logic. Because a single injection provides coverage for a fixed pharmacokinetic interval regardless of the dose or quantity recorded in the dispensing record, the number of days covered is not derived from the dispensed quantity but from the injection interval assigned to that product.

### 4.4.4. Monthly PDC Computation

Rather than computing a single PDC value over the entire observation window, PDC is computed separately for each calendar month. For a given month, the PDC value is the proportion of days in that month that are covered by medication, capped at 1.0. The denominator for the first and last months of the observation window is adjusted to reflect only the days that fall within the window, rather than the full calendar month, avoiding an artificial deflation of coverage at the boundaries.

Within each month, oral and LAI coverage is tracked separately, producing three monthly PDC values per patient: oral, LAI, and total, where the latter represents the union of covered days from both formulation types. This separation is important because it allows the transition from oral to LAI to be characterized at the level of individual months, rather than treating the switch as a clean boundary. In practice, many patients have overlapped oral and LAI coverage in the months immediately surrounding

the switch, and collapsing this into a single metric would obscure the dynamics of the transition period. The final output of this pipeline is a matrix of monthly PDC values for each patient across the 24-month window, which serves as the direct input to the clustering algorithm.

## 4.5. Cohort definition for trajectory analysis

The trajectory analysis cohort was defined through a sequential funnel of inclusion criteria applied to the full PADRIS mental health database. Starting from the total PADRIS mental health cohort of 1,421,510 individuals, the PADRIS-PRESTO study cohort comprises 473,812 individuals with at least one contact with specialized mental health services. From these, 9,969 patients carried a confirmed schizophrenia diagnosis (ICD-10 F20.x), of whom 9,635 had at least one recorded dispensing of an eligible antipsychotic compound. Patients were then restricted to those who switched from oral to LAI treatment between 2015 and 2017, a window justified by the EMA approval landscape: by 2015, the core LAI portfolio was established, with Trevicta receiving European Commission authorization in May 2016, ensuring all switches reflect a contemporary treatment landscape. The 2017 upper bound guarantees at least 12 months of post-switch follow-up within the PADRIS observation period ending in 2019. This yielded 1,147 switchers, of whom 1,025 had complete PDC data across the full 24-month observation window and constituted the final trajectory analysis cohort [Figure 2].

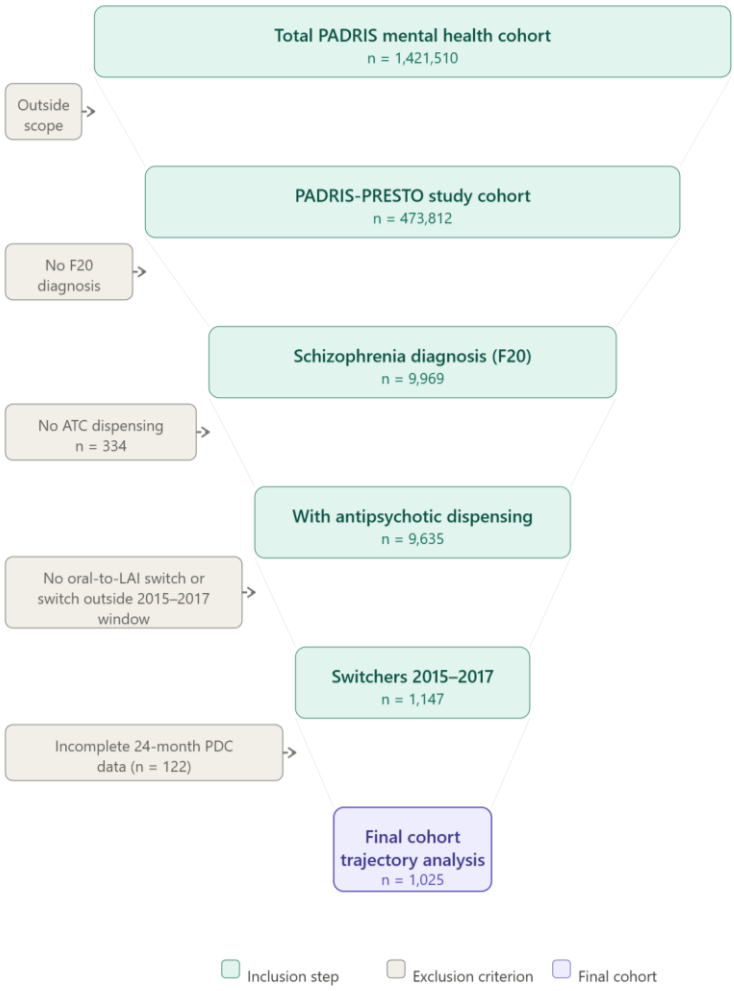



*Figure 2. Stepwise cohort selection from the PADRIS mental health database. to the final trajectory analysis cohort of 1,025 patients with schizophrenia who switched from oral antipsychotics to LAI between 2015 and 2017. Source: own elaboration.*

## 4.6. Time series clustering

The monthly PDC matrix produced in section 4.4 gives each patient a 24-dimensional adherence trajectory. Rather than reducing this to a binary adherent/non-adherent classification, an unsupervised clustering approach was chosen to allow the structure of the data itself to identify distinct longitudinal patterns, without imposing prior assumptions about how many groups exist or what they look like.

The clustering algorithm selected was TimeSeriesKMeans from the tslearn library, using Euclidean distance as the similarity metric. All trajectories were aligned with the same index event, the switch date, so temporal distortion between patients was not expected, making Euclidean distance a suitable and interpretable choice.

The number of clusters k was determined by fitting models across k values from 2 to 10 and evaluating each solution using the silhouette score alongside clinical interpretability, as a statistically optimal k was not sufficient if the resulting clusters did not map onto clinically recognizable adherence profiles, and no group was permitted to represent less than 5% of the total cohort to ensure clinical representativeness of each identified trajectory profile. Cluster structure was additionally validated using UMAP dimensionality reduction from the umap-learn library, which projects the 24-dimensional trajectories into two dimensions. Silhouette scores were computed on this 2D UMAP embedding rather than on the raw trajectory space, as the reduced representation amplifies cluster structure by removing noise dimensions, making it a more sensitive criterion for k selection. The resulting 2D projection was also used for visual inspection of cluster separation.

The full pipeline was implemented in Python using tslearn, pandas, numpy, scipy.stats, pingouin, matplotlib, and seaborn libraries.

## 4.7. Predictive modelling: Identifying candidate patients for LAI switching

The prediction target is the four-class cluster label derived in section 4.6 from the TimeSeriesKMeans solution to the 24-month PDC trajectories of the 1,025 switcher patients. These labels encode four clinically distinct longitudinal adherence profiles identified from real dispensing data and serve as the ground truth for model training. Since cluster assignment requires the full 24-month window, including post-switch data, the objective of the model is to predict this assignment from pre-switch information alone.

It is important to note that this component of the study should be considered exploratory and hypothesis-generating and can be framed as a preliminary proof-of-concept analysis. The model is trained and evaluated on the same cohort from which the cluster labels were derived, and this is not a prospective validation but a methodological exploration of whether pre-switch clinical and dispensing information contains sufficient signal to anticipate post-switch adherence patterns.

The feature set was constructed exclusively from pre-switch information and comprises four categories: the 12 monthly pre-switch PDC values, psychiatric service contacts in the preceding 12 months; sociodemographic variables including sex, age range, smoking status, and socioeconomic level; and pharmacological context variables including oral antipsychotic type and antipsychotic polypharmacy.

Tree-based ensemble methods were selected (Random Forest, XGBoost, and LightGBM) over alternative approaches such as neural networks or logistic regression based on their established performance advantage on tabular clinical datasets, where features are individually meaningful rather than spatially or temporally structured, and for their ability to handle mixed feature types, model non-linear interactions, and produce interpretable feature importance scores. In this setting, gradient-boosted trees and random forests have been shown to consistently outperform both deep learning and linear models [30], while also offering greater interpretability through feature importance methods such as SHAP (SHapley Additive exPlanations) [30]. Hyperparameter optimization was performed using Optuna [31] with stratified 5-fold cross-validation, optimizing macro-averaged AUC across an 80/20 train-test split. SHAP values were computed for the non-PDC features to characterize the direction and magnitude of each feature's influence on cluster assignment, stratified by cluster.

To evaluate the clinical utility of the model beyond the trajectory cohort, the trained classifier was applied prospectively to the 5,662 oral-only patients in the PADRIS dataset who never initiated LAI treatment, using their last 12 months of dispensing data as a proxy for the pre-switch PDC trajectory. All remaining features were constructed using the same pipeline applied to the training cohort.

# 5. Detail Engineering

## 5.1. Synthetic Data Generation

During the administrative approval period preceding access to the real PADRIS data, a synthetic dataset was developed to enable early pipeline development and validation. The generation process was implemented in Python using the pandas and NumPy libraries, producing five CSV files structured to mirror the real PADRIS tables. A pilot dataset of 1,000 individuals was generated first to validate internal consistency, then scaled to 100,000 subjects using extrapolation techniques.

| **Variable** | **Description** | **Frequency / Value** N total (1,421,510) | **Missing values (%)** |
|---|---|---|---|
| Sex n(%) | Women | 53,64% | 0% |
| | Men | 46.36% | 0% |
| Age | mean (sd) | 41,6 (± 22,1) | 0% |
| Age group n(%) | 0–14 | 90,367(9.5%) | 0% |
| | 15–19 | 112,638(11.9%) | |
| | 20–34 | 152,260(16.1%) | |
| | 35–64 | 434,145(45.8%) | |
| | >65 | 158,288(16.7%) | |
| Economical status n(%) | Copayment exempt (Low income/Disability) | 226.208(15,91%) | 15% |
| | < 18.000 € anual | 723.314(50,88%) | |
| | 18.001 - 100.000 € anual | 458.684(32,27%) | |
| | > 100.000 € anual | 13.304 (0,94%) | |
| Birth country n(%) | Europe | 1.285.881(90,5%) | 1% |
| | Spain | 1,094,695(77.5%) | |
| | America | 57.182(4%) | |
| | Africa | 50.698 (3,6%) | |
| | Asia | 26.942 (1,9%) | |
| | Oceania | 78 (<0,1%) | |
| Healthcare region n(%) | Alt Pirineu (Pyrenees and Aran) | 15,538(1.1%) | 2% |
| | Barcelona | 939,319(66.5%) | |
| | Tarragona | 97,463(6.9%) | |
| | Central Catalonia | 113,001(8.0%) | |
| | Girona | 137,013(9.7%) | |
| | Lleida | 79,101(5.6%) | |
| | Terres de l'Ebre | 31,075(2.2%) | |
| Smoking status n(%) | Ever smoker | 586,553(41,51%) | 45,7% |
| | Never smoker | 825,957(58,49%) | |
| BMI (mean, SD) | Body mass index (kg/m$^{2)}$ | 26.96(SD 6.72) | 89,4% |
| Comorbidities count n(%) | 0 | 740,555(52,4%) | 0% |
| | 1 | 317,815(22,5%) | |
| | 2 | 354,540(25,1%) | |
| Comorbidities Specific n(%) | Musculoskeletical disordeers | 605,885(42,88%) | 0% |
| | Hypertension | 263,292(18,64%) | |
| | Arthrosis | 126,137(8,93%) | |

| | | | |
|---|---|---|---|
| | Asthma | 126,137(8,93%) | |
| | Diabetes | 103.961(7,36%) | |
| | Neoplasms | 92,096(6,52%) | |
| | Arthritis | 91,248(6,46%) | |
| | COPD | 69,778(4,94%) | |
| | Osteoporosis | 61,727(4,37%) | |
| | Cronic kidney disease | 47,037(3,33%) | |
| | Ischemic stroke | 38,844(2,75%) | |
| | Heart failure | 33,900(2,40%) | |
| | Dementia | 19,934(2,33%) | |
| | Cirrhosis | 7,486(0,53%) | |
| | AIDS | 7,204(0,51%) | |
| Mes_dispensacio | Dispensing month (YYYYMM) format | | 0% |
| Codi_atc | ATC drug classification code | | 0% |
| Nom_farmac | Drug name | | 0% |
| Num_envasos | Number of packages dispensed ≥ 1 (integer) | | 0% |
| Dies_tractament | Days of treatment covered (DDD) | 4.72 | 5% |
| Preu_pvp | List price in euros | | 0% |
| Aportacio_usuari | Patient co-payment in euros | | 0% |
| Tipus_visita (visits/patient/year) | Outpatient Mental Health Services | 5.36 | 0% |
| | Inpatient Psychiatric Care | 0.78 | 0% |
| | Emergency department | 0.51 | 0% |
| | Emergency psychiatric (Urgències Psiquiàtriques) | 0.43 | |
| Data_ingres | Admission or visit date | | 0% |
| Dies_estada (mean days) | Psychiatric inpatient | 0.37 | 0% |
| | General medical mean | 0.38 | |
| | Emergency-related | 0.28 | |
| Data_alta | Discharge date | | 0% |
| Trastorns Ansietat | Binary flag for anxiety disorders (F40–F48) | 443,225(31.38%) | 0% |
| Trastorns Anim | Binary flag for mood disorders (F30–F39) | 265,976(18.83%) | 0% |
| Esquizofrenia_Psicotics | Binary flag for schizophrenia and psychotic disorders (F20–F29) | 49,438(3.5%) | 0% |
| Trastorns Substancies | Binary flag for substance use disorders | 191,960(13.59%) | 0% |
| Nom_diagnostic | Primary diagnosis label in text format | | 0% |
| Codi_cie10 | ICD-10 diagnostic code (F20-F40) | | 0% |

*Table 3. Structure and descriptive statistics of the synthetic dataset generated to simulate the PADRIS-PRESTO data structure during the administrative approval period **[6] [10] [21]**. Source: own elaboration.*

While the synthetic dataset enabled early development of the full analytical pipeline, it presented inherent limitations. Because it was derived from population-level summary statistics rather than individual-level

records, it could not faithfully reproduce the complex longitudinal patterns of antipsychotic switching, the heterogeneity of LAI injection intervals across products, or the irregularity of real-world dispensing behavior. Furthermore, the psychiatric diagnosis distribution was simplified relative to the real PADRIS population, and comorbidity patterns were assigned independently rather than reflecting the clinical correlations that exist in real patient data. For these reasons, all substantive analyses were conducted exclusively on the real PADRIS dataset once it became available.

## 5.2. Analytical Pipeline

### 5.2.1. Data Sources and Tables

Table 23 (farmacs_dispensats) contains all community pharmacy dispensing events. Each row represents one prescription dispensed to one patient in one month. The key columns are: idCas (anonymous patient ID), Any_Mes (year-month in YYYYMM format), DES_HIS_Producte (commercial drug name), and Nombre_DDD_Receptes_Dispensades (the total number of Defined Daily Doses dispensed, which translates directly to days of coverage for oral medications).

Table 25 (farmacs_mhda) contains hospital-dispensed medications, which in this context are almost exclusively LAI antipsychotics administered under clinical supervision. This table does not have a DDD column, so coverage for LAIs must be derived from the known injection interval of each product.

Three tables record psychiatric contacts: URG (emergency department visits), SMH (psychiatric hospitalisations), and SMA (outpatient mental health consultations). Each contains patient IDs, admission and discharge dates, and ICD diagnosis codes across four diagnostic positions.

The cohort_schizophrenia_sociodemografica table contains one row per patient and includes sex, age range, nationality, socioeconomic level, BMI (first and last recorded), and smoking status variables (never_smoker, current_smoker, ex_smoker, and a VAL_change variable indicating whether smoking status changed over the observation period).

### 5.2.3. Drug Formulation Classification

Before calculating adherence, every dispensing event must be labelled as one of three types: ORAL, LAI, or IMMEDIATE_ACTION. This classification is based on the commercial product name stored in the dispensing tables [Table 4].

Each LAI product was matched to its approved injection interval using the Summary of Product Characteristics (SmPC) from the Spanish medicines agency (CIMA) or the European Medicines Agency. A Python dictionary maps each product name to its interval in days:

| Product | Active Substance | Interval |
|---|---|---|
| Xeplion (all doses) | Paliperidone palmitate 1-month | 28 days |
| Abilify Maintena 300mg, 400mg | Arpiprazole 1-month | 28 days |
| Trevicta (all doses) | Paliperidone palmitate 3-month | 84 days |
| Risperidal Consta (all doses) | Risperidone | 14 days |
| Zypadhera 210mg | Olanzapine pamoate | 14 days |
| Zypadhera 300mg, 405mg | Olanzapine pamoate | 28 days (conservative) |

| | | |
|---|---|---|
| Haldol Decanoate / Depot | Haloperidol decanoate | 28 days |

*Table 4. LAI antipsychotic products included in the study with their corresponding active substance and assigned injection interval, as specified in the Summary of Product Characteristics (SmPC) registered in CIMA and the EMA.*

Immediate-action injectables (short-acting, used for acute agitation) are excluded from PDC calculations but kept in the dataset for potential future analyses. These include ZYPREXA 10mg vials, ABILIFY 7.5mg vials, DOGMATIL ampoules, and HALOPERIDOL ESTEVE ampoules. All remaining formulations are classified as ORAL. The classify_product() function applies this logic: it checks whether a product name matches any key in the LAI dictionary, then checks the IMMEDIATE_ACTION list, and otherwise labels the row as ORAL.

### 5.2.4. PDC Calculation

The Proportion of Days Covered (PDC) is the primary adherence metric. For a given time period (month or year), PDC equals the days the patient had medication available divided by the days in that period. A value of 1.0 means the patient was covered every day; 0.0 means no coverage at all.

The first challenge is that Any_Mes only records the year and month of dispensing, not the exact date. We assume dispensing occurred on the 15th of each month as a reasonable midpoint. This is implemented by the anymes_to_date() function, which parses the YYYYMM integer and returns a Python Timestamp for the 15th of that month.

For oral medications, the number of days covered is taken directly from Nombre_DDD_Receptes_Dispensades, which already represents the total days of treatment **across all dispensed units**. For example, if a patient receives 2 boxes of a 30-DDD medication, this column contains 60. For LAI medications, the days covered are fixed by the injection interval. Receiving one XEPLION injection means the patient is covered for exactly 28 days from the injection date, regardless of how many units were dispensed.

A critical methodological decision is how to handle early refills, which are situations where a patient picks up a new prescription before the previous one runs out. The standard PDC approach is to carry the leftover days forward: the new coverage period begins only after the previous one ends. This prevents double-counting and more accurately reflects actual drug availability. The merge_intervals() function implements this. It takes a list of (start_date, end_date) coverage intervals, sorts them chronologically, and whenever two intervals overlap, it shifts the new interval forward by the number of leftover days. The result is a list of non-overlapping intervals that represent the patient's total coverage timeline.

Two PDC resolutions are computed. Annual PDC divides the total covered days within a calendar year by the number of days in that year (adjusted in the first year of treatment to start from the treatment start date, not January 1st). Monthly PDC uses the same logic but with a denominator equal to the number of days in each specific month (28, 29, 30, or 31). Three PDC values are calculated per patient per period: PDC_oral (from oral dispensings only), PDC_lai (from LAI dispensings only), and PDC_total (the union of both, so a day covered by either type counts once). All PDC values are capped at 1.0. The compute_pdc() and compute_monthly_pdc() functions implement this logic. They take a patient's dispensing data from both tables, build separate coverage interval lists for oral and LAI, merge each list, and then call days_covered_in_year() or days_covered_in_month() to count covered days within each time window.

The treatment start date (first dispensing event across both tables) is computed once per patient and used as the denominator adjustment anchor for the first year only. All subsequent years use the full 365/366 days as the denominator.

### 5.2.5. Cohort Definition and Patient Selection

The trajectory analysis was restricted to patients in the study cohort, excluding the control population. Diagnosis was identified using ICD-10 codes starting with F20 in the diagnosis table, capturing all schizophrenia spectrum presentations [Table 1]. From the 473,812 individuals in the study cohort, 9,969 carried a confirmed schizophrenia diagnosis. Of these, 334 had no recorded dispensing of any of the eleven antipsychotic compounds included in the ATC selection [Table 4] and were therefore excluded, leaving 9,635 patients with at least one eligible antipsychotic dispensing event.

The trajectory analysis required patients who transitioned from oral antipsychotic treatment to LAI. A patient was classified as a switcher if they had oral antipsychotic dispensing events recorded before their first LAI dispensing event. The index date, named switch date, defined as the date of the first LAI dispensing event per patient, served as the anchor point around which the entire 24-month observation window was constructed. This definition carries an important implication as it is based on dispensing records rather than prescribing records. A patient who was prescribed LAI but for whom no dispensing was recorded would not be captured as a switcher.

To be included in the trajectory cohort, a patient's index date was required to fall between January 2015 and December 2017. The 2015 lower bound reflects the EMA approval landscape: by 2015, the core portfolio of LAI formulations included in this study had received European Medicines Agency approval, with the last addition being paliperidone palmitate 3-month injection (Trevicta), which received European Commission authorization in May 2016 [32], ensuring that switches occurring from 2015 onwards reflect a contemporary and complete treatment landscape. The 2017 upper bound ensures that every patient in the cohort has at least 12 months of post-switch follow-up data available.

This yielded 1,147 switchers within the window. Only patients with complete PDC data 12 months before and 12 months after the switching date were retained. Of the 1,147 switchers, 122 were excluded due to incomplete previous or posterior year of PDC data, yielding a final trajectory analysis cohort of 1,025 patients with complete longitudinal data across the full 24-month window centered on their LAI initiation date [Fig. 1].

### 5.2.6. Trajectory Clustering with TimeSeriesKMeans

Rather than summarizing adherence as a single number per patient, the evolution of adherence over the 24-month window was modeled to identify distinct trajectory subgroups. The monthly PDC values were arranged into a matrix of shape (1025, 24), where each row corresponds to a patient and each column to one of the 24 months (Month_-12 through Month_-1 for the pre-switch period, Month_1 through Month_12 for the post-switch period). This matrix was reshaped into a three-dimensional array shape (1025, 24, 1) as required by the tslearn library, where the third dimension holds the single PDC value for each month.

To determine the optimal number of trajectory clusters, a two-stage approach was adopted. In the first stage, UMAP (Uniform Manifold Approximation and Projection) from the umap-learn library was applied to reduce the 24-dimensional PDC trajectory matrix to a 2-dimensional embedding, preserving the

underlying structure of the data while making it interpretable visually. UMAP works by constructing a high-dimensional graph of neighborhood relationships between data points and then optimizing a low-dimensional representation that preserves this structure as faithfully as possible.

In the second stage, k-means clustering was applied directly to the 2D UMAP embedding for k values ranging from 2 to 10. For each value of k, the silhouette score was computed. The silhouette score for a given patient is defined as $\frac{(b-a)}{\max(a,b)}$ where a is the mean distance to all other patients in the same cluster and b is the mean distance to all patients in the nearest neighboring cluster. Values close to 1 indicate that the patient is well-matched to their own cluster and clearly separated from others, while values close to 0 indicate overlap between clusters and values below 0 suggest misassignment. The mean silhouette score across all patients was used as the cluster quality metric for each k.

| k value | Silhouette score (0-1) |
|---|---|
| 2 | 0.5395 |
| 3 | 0.5299 |
| 4 | 0.5505 |
| 5 | 0.5078 |
| 6 | 0.5144 |
| 7 | 0.4940 |
| 8 | 0.4721 |
| 9 | 0.4747 |
| 10 | 0.4793 |

*Table 5. Silhouette scores obtained for TimeSeriesKMeans solutions evaluated across k=2 to k=10 on the 2D UMAP embedding of the 24-month PDC trajectory matrix. Yellow highlight indicates the selected solution. Source: own elaboration.*

The highest silhouette score was obtained at k=4 (0.5505) [Table 5], indicating that four clusters provided the best balance between within-cluster cohesion and between-cluster separation.
While the silhouette score alone would justify k=4 as the optimal solution, clinical interpretability provided an additional layer of validation. The k=3 solution was rejected because it failed to separate the early dropout subgroup from the low-adherence improvers, collapsing two clinically distinct profiles into a single cluster. The k=5 solution produced a fifth cluster that was not clinically distinguishable from an existing one and showed a lower silhouette score. The k=4 solution was therefore the only one that simultaneously satisfied the quantitative criterion and produced four profiles that map onto recognizable and distinct clinical trajectories. An additional minimum cluster size criterion was applied at this stage, requiring that no cluster represent less than 5% of the total cohort (approximately 51 patients out of 1,025) to ensure that each identified trajectory profile was clinically meaningful and not driven by a small outlier subgroup. All four clusters in the final k=4 solution satisfied this criterion, with the smallest group comprising 187 patients (18.2% of the cohort) [Fig. 3].
Importantly, this result was obtained on the UMAP-reduced space rather than on the raw 24-dimensional trajectories, which explains why the absolute silhouette values are higher than those computed directly on the PDC matrix (where the maximum was 0.272 at k=2).

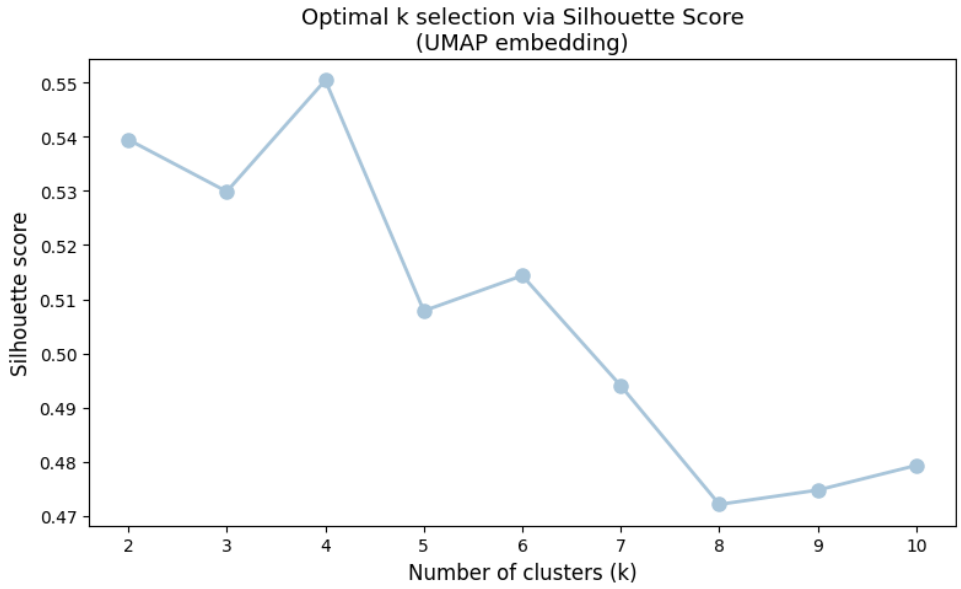



*Figure 3. Silhouette score as a function of the number of clusters k, computed on the 2D UMAP mbedding of the 24-month PDC trajectories for k values ranging from 2 to 10. The peak k=4 (silhouette = 0.5505) guided the selection of the final clustering solution. Values above the x-axis represent the mean silhouette score across all 1,025 patients for each k.*

The k=4 solution was therefore validated both statistically, through the silhouette criterion on the UMAP embedding, and visually, through inspection of the 2D UMAP projection [Fig. 4] where the four groups appear as spatially distinct regions.

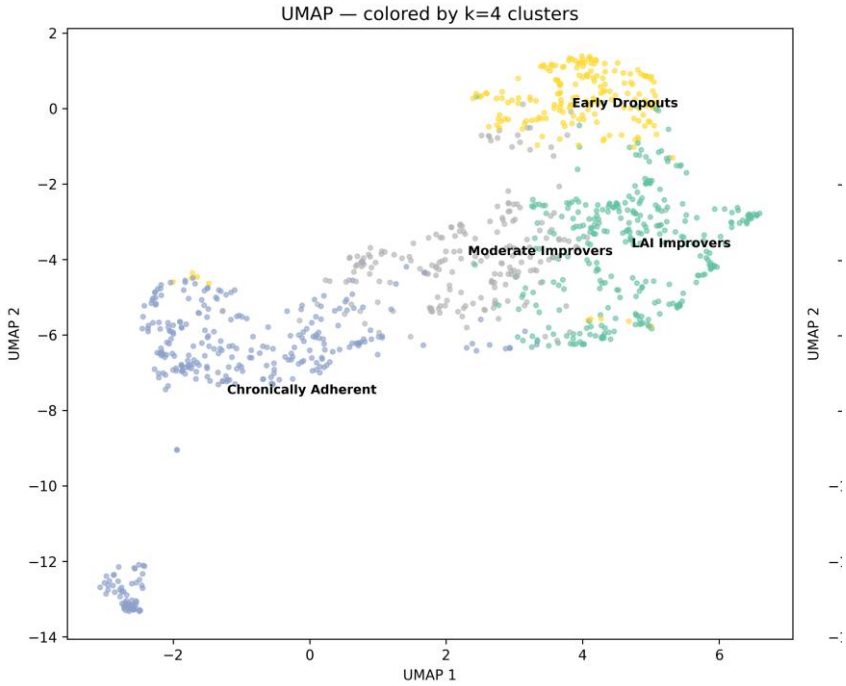



*Figure 4. 2D UMAP projection of the 24-month PDC trajectories for the 1,025 patients in the trajectory analysis cohort, colored by cluster assignment from the k=4 TimeSeriesKMeans solution. Spatial separation between the four groups provides visual confirmation of the clustering structure identified through the silhouette criterion. Source: own elaboration.*

Having determined k=4 as the optimal solution, TimeSeriesKMeans from the tslearn library was fitted on the original 24-dimensional PDC trajectories using Euclidean distance as the similarity metric, with random state fixed at 42 for reproducibility.

Both continuous PDC (values 0–1) and binary PDC (1 if PDC ≥ 80%, else 0) were tested. Continuous PDC was chosen [Table 6] because it preserves the full degree of adherence information rather than collapsing it into a dichotomy, and because it produced equivalent silhouette scores to the binary version, meaning that no discriminative power was gained by binarizing while information was lost.

| idCas | Pre and post-switch months | | | | | | | | cluster_k4 |
|---|---|---|---|---|---|---|---|---|---|
| | -4 | -3 | -2 | -1 | 1 | 2 | 3 | 4 | |
| 83951106 | 0.467 | 0.9 | 0.9 | 1.0 | 0.933 | 0.933 | 0.933 | 0.933 | 3 |
| 83996724 | 0.0 | 0.0 | 0.033 | 0.57 | 0.933 | 0.6 | 0.0 | 0.0 | 2 |
| 84015456 | 0.0 | 0.033 | 0.367 | 1.0 | 0.933 | 0.967 | 0.967 | 0.933 | 0 |
| 84051362 | 0.967 | 0.967 | 0.0 | 0.93 | 0.933 | 0.933 | 0.933 | 0.933 | 0 |

*Table 6. Sample of the first ten patients in the monthly PDC matrix showing months -4 to +4 relative to the switch date, alongside their assigned trajectory cluster (k=4 solution). PDC values range from 0 to 1, where 1 indicates full medication coverage for that month. Month 0 (the switch month) is excluded from the observation window because it represents a transition period during which represents a transition period where oral and LAI coverage may overlap. Source: own elaboration.*

### 5.2.7. Adherence Trajectory Clusters and Healthcare Contact Patterns

Before examining the individual cluster profiles, the overall adherence patterns across the full trajectory cohort are presented to contextualize the clustering results. Figure 5 shows the individual PDC trajectories for a random sample of 100 patients alongside the cohort mean, revealing substantial inter-patient variability in pre-switch oral adherence and a clear population-level improvement following LAI initiation, with the mean post-switch PDC rising from approximately 50% to 70-90%. A random sample of 100 patients was selected for visualization rather than the full cohort of 1,025, as plotting all individual trajectories simultaneously would produce an uninterpretable figure due to severe overplotting.

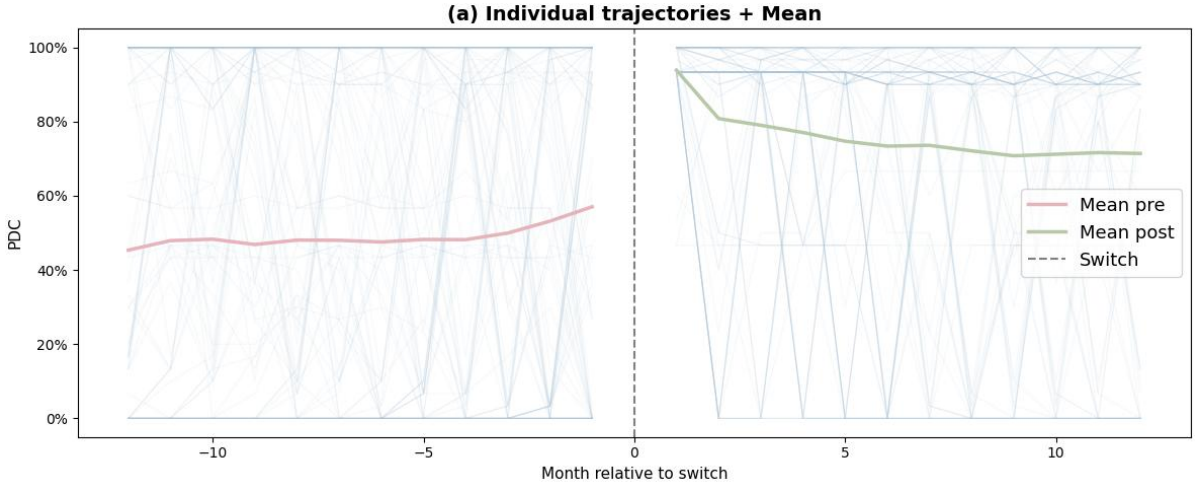



*Figure 5. Individual PDC trajectories for a random sample of 100 patients (thin blue lines), overlaid with the cohort mean pre-switch (pink) and post-switch (green). The dashed vertical line indicates the switch date (month 0). Source: own elaboration.*

Figure 6 reinforces this pattern at the patient level: the heatmap of 200 randomly sampled patients shows a predominantly heterogeneous pre-switch adherence landscape that transitions to a predominantly green post-switch pattern, though a subset of patients remains poorly adherent throughout. The pre-switch period shows heterogeneous adherence, while the post-switch period shifts predominantly towards full coverage, with a visible subset of patients remaining poorly adherent.

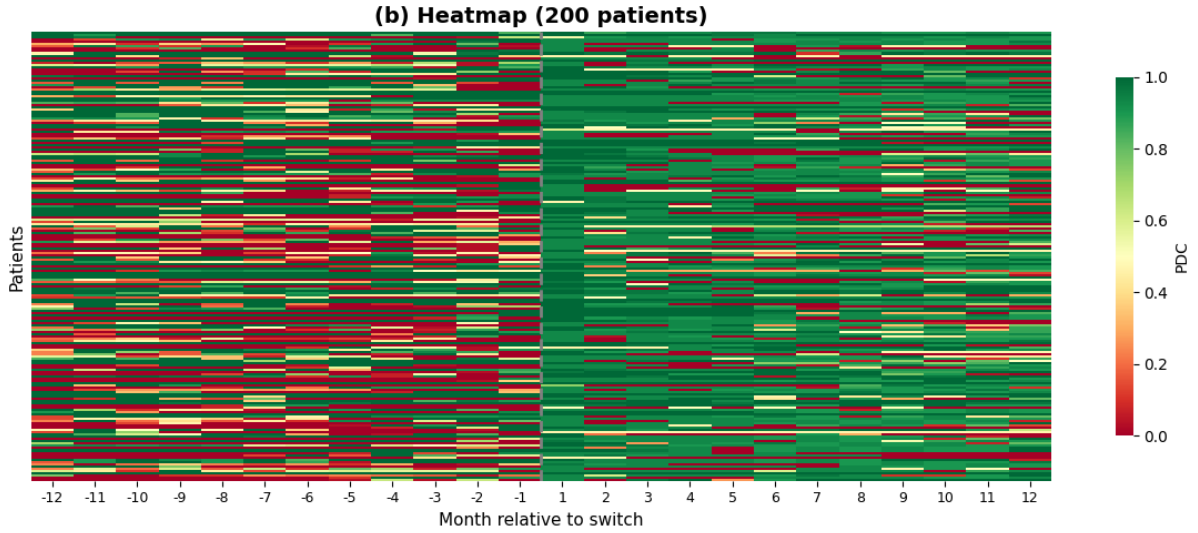



*Figure 6. Heatmap of monthly PDC values for a random sample of 200 patients across the full 24-month observation window. Each row represents one patient and each column one month, with color encoding PDC from 0 (red) to 1 (green). The dashed vertical line indicates the switch date. Source: own elaboration.*

Figure 7 illustrates the distributional shift more formally: pre-switch PDC distributions are wide and bimodal, with patients concentrated at both 0% and 100%, while post-switch distributions narrow sharply towards full coverage.

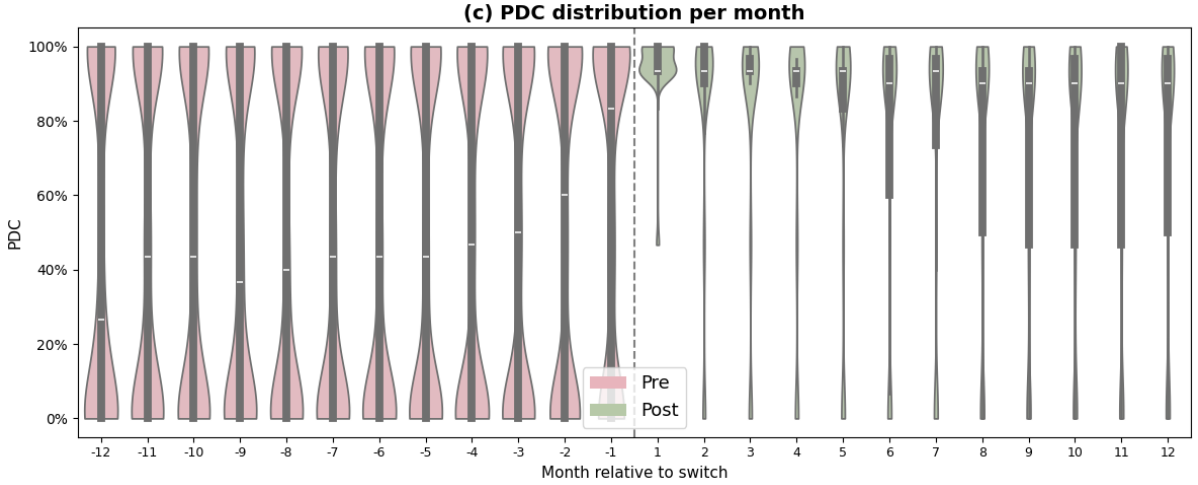



*Figure 7. Violin plots of the monthly PDC distribution or each of the 24 observation months, colored by period (pre-switch in pink, post-switch in green). Source: own elaboration.*

These population-level observations motivate the trajectory-based approach: a single summary statistic would obscure the heterogeneity visible in these figures.

The k=4 TimeSeriesKMeans solution identified four clinically distinct adherence trajectory profiles across the 1,025 patients in the trajectory analysis cohort. The sociodemographic and clinical characteristics of each group are summarized in Table 7, confirming that the identified clusters differ not only in their longitudinal adherence patterns but also in key clinical and demographic features. Each cluster is characterized by its pre-switch oral adherence pattern and its post-switch LAI adherence response, together defining a unique clinical profile with distinct implications for treatment management. The four profiles are presented below alongside their individual trajectory plots, with the 80% PDC threshold indicated as a clinical reference point.

To assess the clinical impact of adherence trajectories, psychiatric contacts were linked to the adherence data. Three visit types were analyzed: emergency department visits (URG), psychiatric hospitalizations (SMH), and outpatient mental health consultations (SMA). All three tables were filtered to the same ±12-month window around each patient's switch date.

Prior to the quantitative analysis of healthcare contacts, the psychiatric relevance of recorded episodes was formally validated. In the URG table, 81.0% of episodes carried a psychiatric primary diagnosis, rising to 83.9% when the secondary diagnosis was also considered. In the SMH table, the corresponding figures were 88.2% and 89.9%. The most frequent primary diagnoses in the URG table included unspecified psychosis, paranoid schizophrenia subtypes, and schizoaffective disorder. In the SMH table, paranoid schizophrenia with acute exacerbation, unspecified psychosis, and schizoaffective disorder accounted for the majority of admissions. In the SMA table, 97.8% of episodes had a psychiatric primary diagnosis, rising to 98.1% when including the secondary position, the highest validation rate across the three tables, consistent with the outpatient mental health setting where non-psychiatric contacts would be exceptional. The SMA table, which stores ICD-9 codes, showed paranoid schizophrenia (295.30), paranoid schizophrenia with acute exacerbation (295.32), and unspecified psychosis (298.9) as the most frequent diagnoses. These findings confirm that the recorded contacts reflect genuine psychiatric episodes relevant to the study population, supporting the validity of using these contact counts as clinical outcome measures.

| Cluster name | Cluster 0 (LAI Improvers) | Cluster 1 (Chronically Adherent) | Cluster 2 (Early Dropouts) | Cluster 3 (Moderate Improvers) |
|---|---|---|---|---|
| N | 321 | 319 | 187 | 198 |
| Male (%) | 30.8% | 27.9% | 31.0% | 34.8% |
| Median age range | 35 - 39 | 40 - 44 | 40 - 44 | 40 - 44 |
| NIVSOC 1. Exempts | 147 (45.8%) | 163 (51.1%) | 81 (43.3%) | 92 (46.5%) |
| NIVSOC 2. < 18.000 | 154 (48.0%) | 145 (45.5%) | 101 (54.0%) | 101 (51.0%) |
| NIVSOC 3. 18.001 - 100.000 | 18 (5.6%) | 10 (3.1%) | 5 (2.7%) | 5 (2.5%) |
| NIVSOC 4. > 100.000 | 2 (0.6%) | 1 (0.3%) | 0 (0.0%) | 0 (0.0%) |
| Never smoker (%) | 27.2% | 19.9% | 38.4% | 20.8% |
| Current smoker (%) | 72.4% | 80.1% | 60.9% | 79.2% |
| Ex-smoker (%) | 0.4% | 0.0% | 0.7% | 0.0% |
| Main oral type | aripiprazole | aripiprazole | aripiprazole | aripiprazole |
| Main LAI type | paliperidone | paliperidone | paliperidone | paliperidone |
| Main switch year | 2016 | 2015 | 2015 | 2015 |
| Antipsychotic polypharmacy | 2.0 (2.0-3.0) | 3.0 (2.0-4.0) | 3.0 (2.0-3.0) | 3.0 (2.0-4.0) |

*Table 7. Sociodemographic and clinical characteristics of the four adherence trajectory clusters identified by the k=4 TimeSeriesKMeans solution among the 1,025 patients in the trajectory analysis cohort. NIVSOC: socioeconomic level categories defined by the Catalan Health System. Antipsychotic polypharmacy is expressed as median (IQR). Main oral type reflects the most frequent named active substance dispensed in the pre-switch year, excluding unclassified products. Main LAI type reflects the most frequent LAI product initiated at the switch date. Source: own elaboration.*

For each month relative to the switch date, the number of visits per cluster is shown as stacked bars (colored by visit type) overlaid on the mean PDC line. This combined visualization allows the reader to see whether reductions in adherence correspond temporally to increases in clinical contacts.

- **Cluster 0: LAI Improvers** (n=321, 31.3%)

This cluster represents the patients who benefit most from switching to LAI. Pre-switch oral adherence was consistently very low throughout the entire 12-month pre-switch period, with a mean PDC below 10% in the early months, rising gradually to approximately 75% in the final pre-switch month, suggesting an anticipated or preparation effect around the time of the switch decision [Fig. 8]. Following LAI initiation, adherence rises sharply, peaking at approximately 95% in month 1 before stabilizing around the 80% adherence threshold for the remainder of the post-switch year. The narrow 95% confidence interval post-switch indicates consistent behavior across patients in this group. This is the clinically most relevant cluster, representing patients for whom LAI switching produced a substantial and sustained adherence benefit.

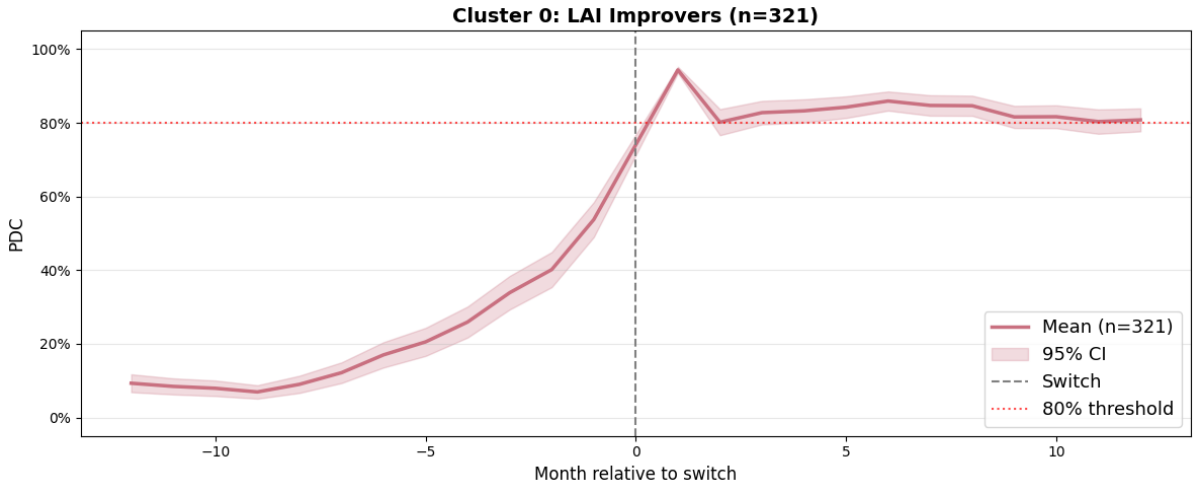



*Figure 8. Mean PDC trajectory with 95% confidence interval for Cluster 0 (LAI Improvers, n=321). The dashed vertical line indicates the switch date and the dotted red line for the 80% adherence threshold. Source: own elaboration.*

The progressive rise in adherence in the final pre-switch months, reaching approximately 75% by month -1, coincides with a marked increase in psychiatric contacts across all three service types [Fig.9], suggesting that intensified clinical engagement in the period preceding the switch decision may have contributed to improved oral adherence and ultimately motivated the transition to LAI. Due to the switch, psychiatric contacts decrease substantially, with emergency and inpatient contacts showing the most pronounced reductions. This parallel improvement in adherence and reduction in healthcare utilization is consistent with the hypothesis that LAI initiation produces clinical stabilization in patients who were previously non-adherent.

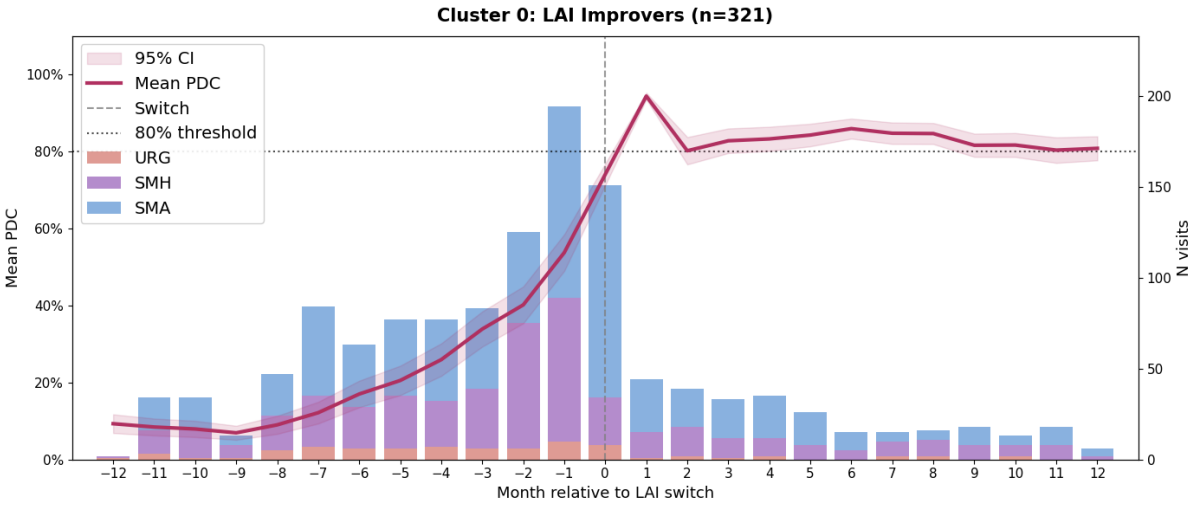



*Figure 9. Mean PDC trajectory with 95% confidence interval (left axis) and monthly psychiatric contact counts of Cluster 0 by service type (right axis, stacked bars). URG: emergency department; SMH: inpatient psychiatric care; SMA: outpatient mental health services. Source: own elaboration.*

- **Cluster 1: Chronically Adherent** (n=319, 31.1%)

This cluster comprises patients who were already highly adherent to oral antipsychotics before switching and remained so after. The mean PDC remained consistently above 80% throughout the entire 24-month window, with minimal variability and an extremely narrow confidence interval, indicating a homogeneous and stable adherence profile [Fig. 10]. The switch to LAI did not produce a notable change in adherence level, which is expected: these patients had no adherence deficit to correct.

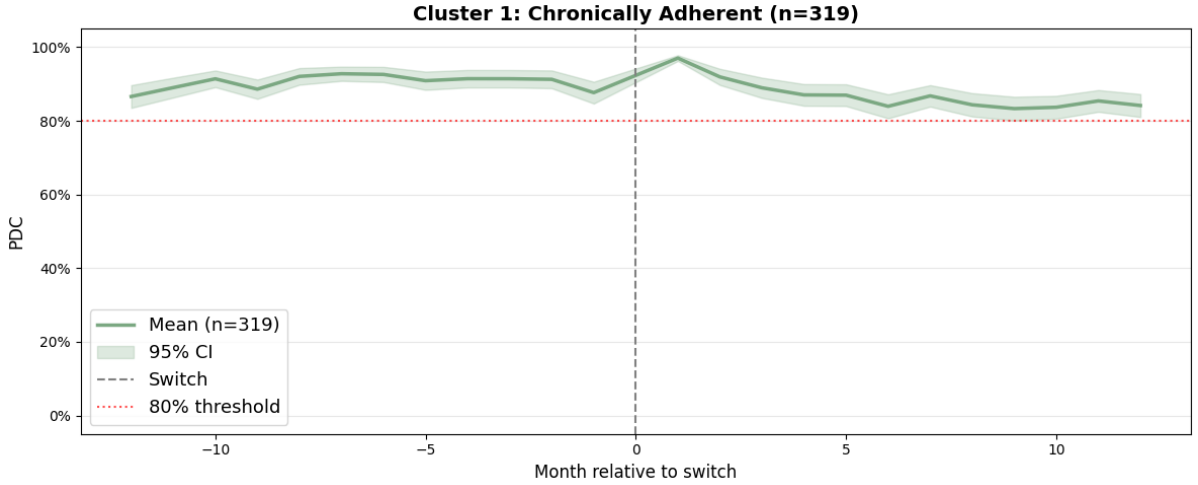



*Figure 10. . Mean PDC trajectory with 95% confidence interval for Cluster 1 (Chronically Adherent, n=319). Individual trajectories for a random sample of 50 patients are shown in the background. The dashed vertical line indicates the switch date and the dotted red line the 80% adherence threshold. Source: own elaboration.*

Psychiatric contacts are distributed across the full observation window without a marked pre-switch peak, reflecting a pattern of stable ongoing clinical follow-up rather than acute decompensation. Post-switch contacts show a modest reduction but remain relatively stable [Fig.11], suggesting that this group maintains regular engagement with mental health services regardless of treatment modality. This raises an important clinical consideration: while LAI switching did not improve adherence in this group, it may still offer advantages in terms of guaranteed administration continuity and reduced relapse risk during periods of voluntary disengagement.

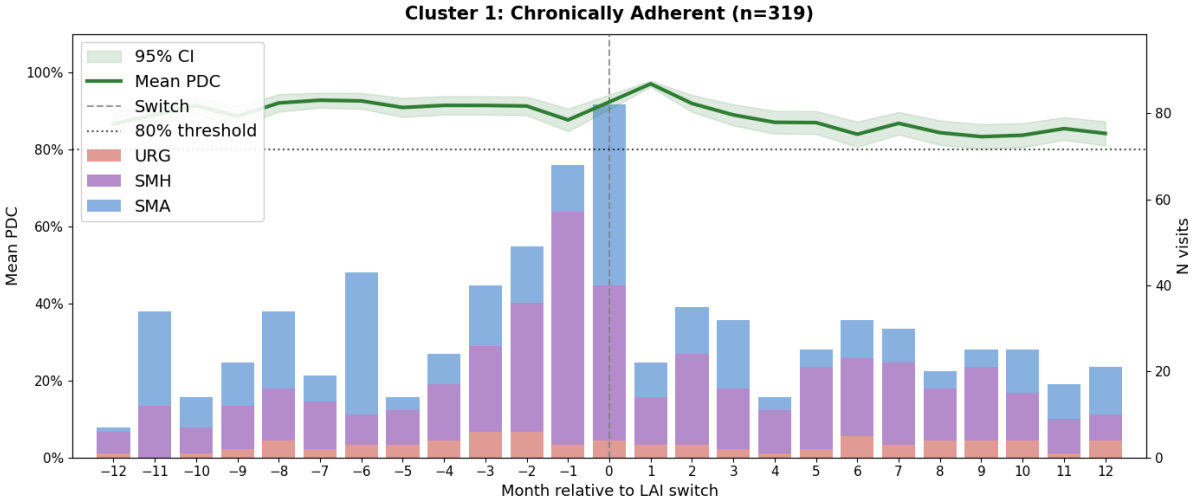



*Figure 11. Mean PDC trajectory with 95% confidence interval (left axis) and monthly psychiatric contact for Cluster 1 total counts by service type (right axis, stacked bars) for Cluster 1, Chronically Adherent (n=319). Source: own elaboration.*

- **Cluster 2: Early Dropouts** (n=187, 18.2%)

This cluster shows the most clinically concerning trajectory. Pre-switch oral adherence was low but stable, around 20-30% throughout the pre-switch year [Fig.12], with a notable increase in the two months immediately preceding the switch, possibly reflecting a period of intensified clinical engagement. Following LAI initiation, adherence spikes sharply to approximately 90% in month 1, briefly crossing the 80% threshold, before declining rapidly and continuously to below 20% by months 7-8, where it remains for the rest of the observation period. This pattern is consistent with early LAI discontinuation patients who initiated treatment but failed to maintain it. The wide confidence interval in the post-switch period reflects substantial heterogeneity in the timing and rate of dropout across patients in this group.

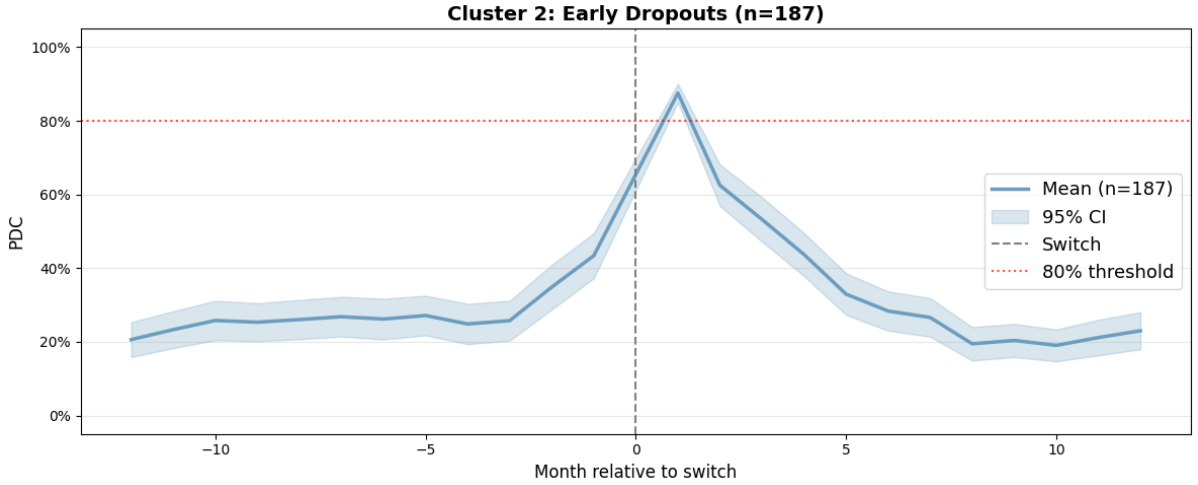



*Figure 12. Mean PDC trajectory with 95% confidence interval for Cluster 2 (Early Dropouts, n=187). Individual trajectories for a random sample of 50 patients are shown in the background. The dashed vertical line indicates the switch date and the dotted red line for the 80% adherence threshold. Source: own elaboration.*

The high volume of pre-switch inpatient contacts, visible as the dominant component of the stacked bars in the final pre-switch months, reflects the clinical severity of this group. Post-switch, psychiatric contacts remain elevated [Fig. 13] relative to other clusters, consistent with ongoing clinical instability. This is the most clinically concerning trajectory profile: patients who initiated LAI treatment but failed to sustain it, and who continued to require high levels of psychiatric care.

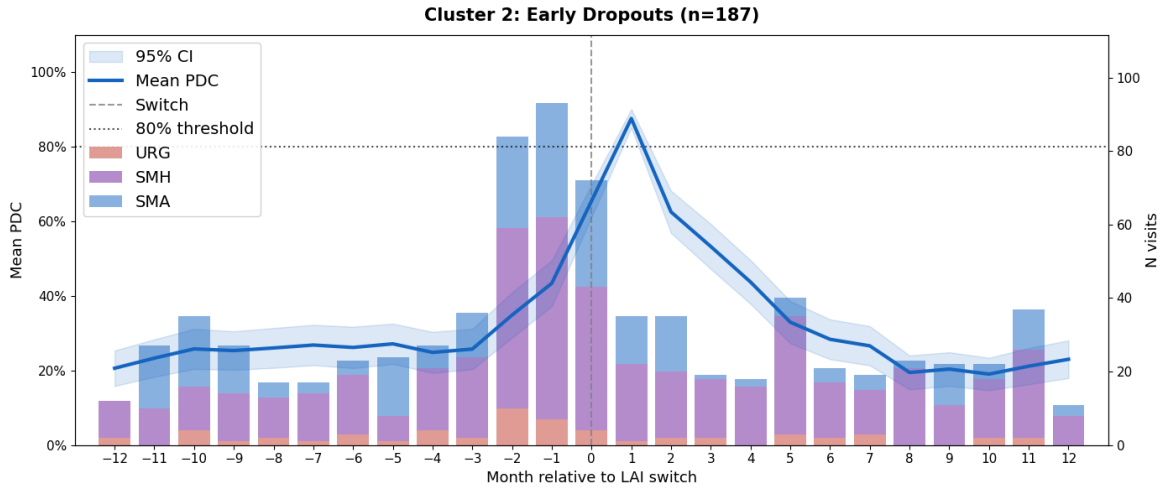



*Figure 13. Figure 12. Mean PDC trajectory with 95% confidence interval (left axis) and monthly psychiatric contact total counts for Cluster 2 by service type (right axis, stacked bars). Source: own elaboration.*

- **Cluster 3: Moderate Improvers** (n=198, 19.3%)

This cluster shows a partial adherence improvement following the switch. Pre-switch oral adherence was moderate, starting around 60-70% in the early pre-switch months and declining progressively to approximately 25% [Fig. 14] in the month immediately before the switch, suggesting a deterioration in oral adherence that may have motivated the switch decision. Following LAI initiation, adherence rises sharply to approximately 95% in month 1, then stabilizes around the 80% threshold for the remainder of the post-switch year, closely mirroring the trajectory of the LAI Improvers but not completely passing the 80% threshold every month. The mean post-switch PDC hovers at the 80% boundary throughout, making this group the most clinically ambiguous. These are patients who achieve the adherence threshold but with little margin, and whose long-term trajectory remains uncertain.

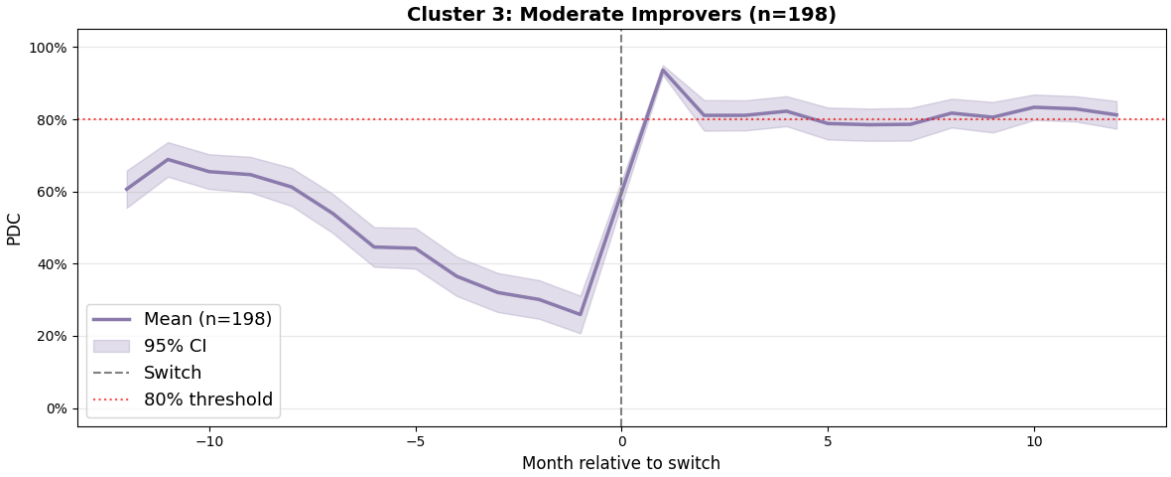



*Figure 14. Mean PDC trajectory with 95% confidence interval for Cluster 3 The dashed vertical line indicates the switch date and the dotted red line the 80% adherence threshold. Source: own elaboration.*

It is worth noting that during the cluster selection process, a profile of patients who were highly adherent to oral antipsychotics but showed a deterioration in adherence following LAI initiation was anticipated as a clinically plausible trajectory. No such cluster emerged from the data, suggesting that in this cohort, switching to LAI did not produce sustained adherence worsening in previously adherent patients. The absence of this profile is itself a meaningful finding, as it provides population-level evidence against the hypothesis that LAI switching is systematically detrimental to adherent patients.

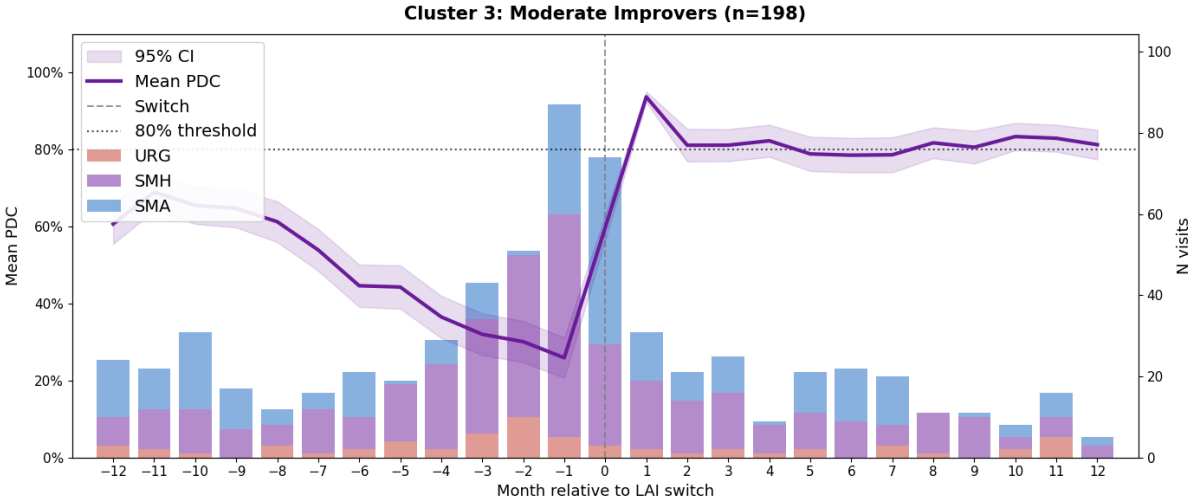
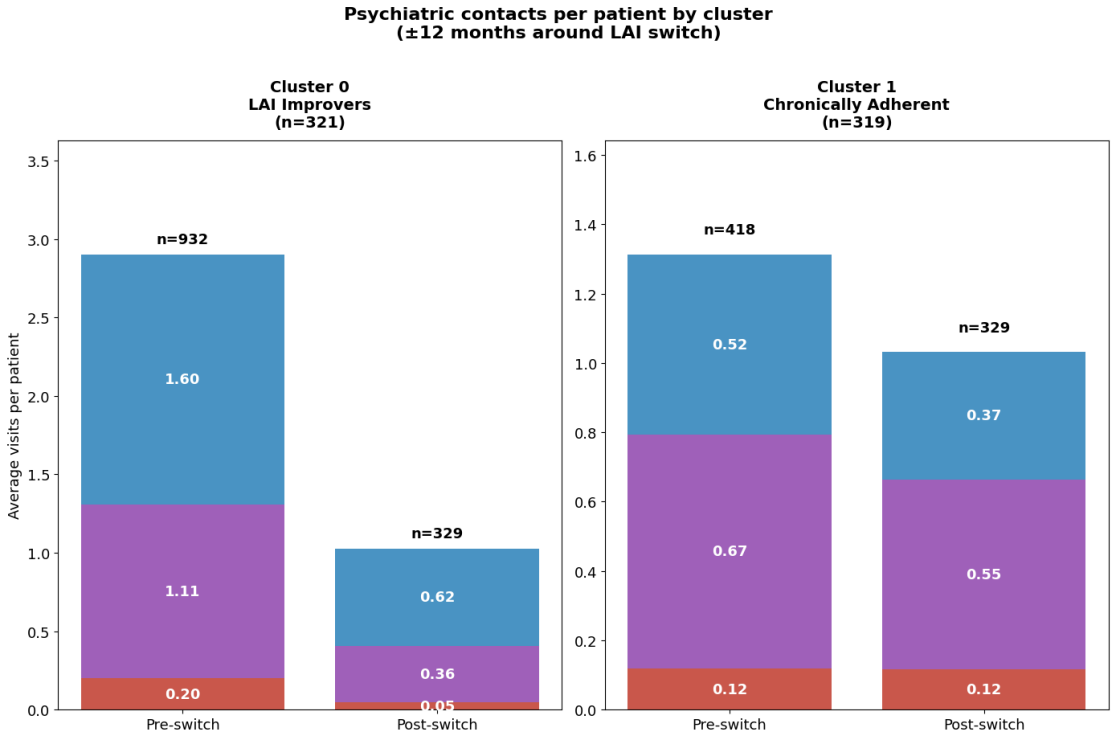



*Figure 15. Mean PDC trajectory with 95% confidence interval (left axis) and monthly total psychiatric contact counts for Cluster 3 by service type (right axis, stacked bars). Source: own elaboration.*

Psychiatric contacts show a clear and sustained reduction of visits post-switch, particularly in inpatient admissions, suggesting clinical stabilization despite the borderline adherence level [Fig. 15].

The stacked bar charts [Fig. 16] present the average number of psychiatric contacts per patient in the 12 months before and after the switch, disaggregated by service type across the four clusters. Cluster 0 (LAI Improvers) showed the largest absolute reduction in total contacts, from 932 pre-switch to 329 post-switch, driven primarily by a reduction in outpatient mental health visits (SMA: 1.60 to 0.62). Cluster 1 (Chronically Adherent) showed a modest reduction from 418 to 329, with proportionally stable contributions from all three service types, consistent with the pattern of stable ongoing follow-up rather than acute decompensation.



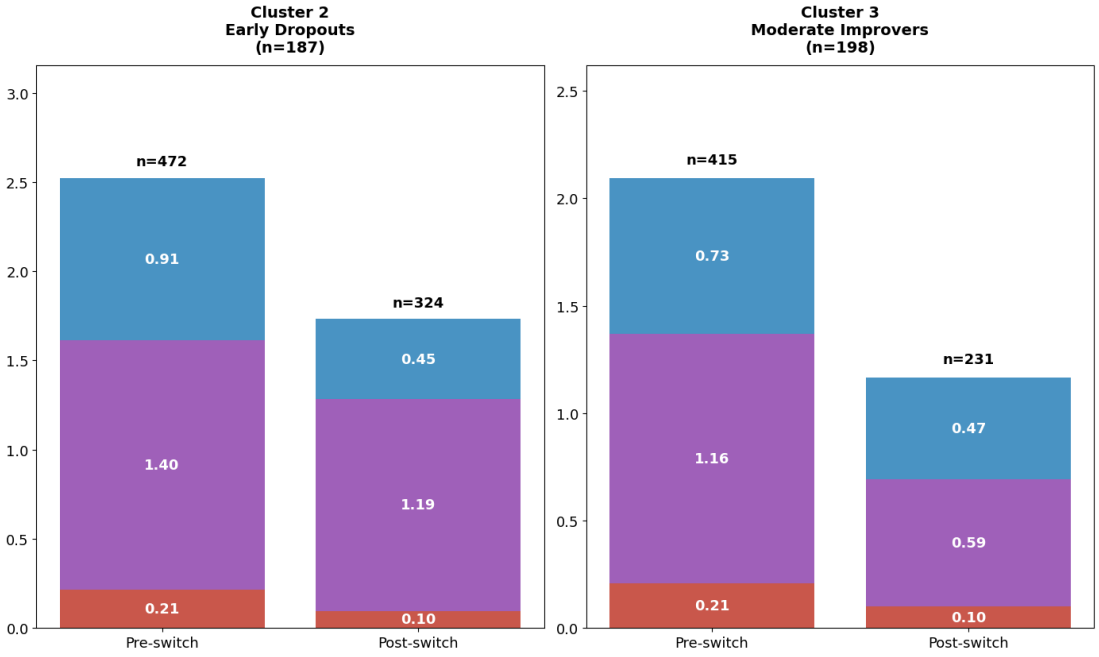



*Figure 16. Summary comparison of average psychiatric contacts per patient across the four clusters , split by pre- and post-switch period. Each stacked bar shows the contribution of URG, SMH, and SMA visits; total visit counts (n) are labelled at the top.*

Cluster 2 (Early Dropouts) showed the smallest reduction in total contacts from total visits 472 to 324, with inpatient admissions remaining persistently elevated post-switch (SMH: 1.40 to 1.19), reflecting continued clinical instability despite LAI initiation. Cluster 3 (Moderate Improvers) showed a reduction from 415 to 231 [Fig. 16], with inpatient contacts showing the largest absolute decrease (SMH: 1.16 to 0.59), suggesting clinical stabilization consistent with the adherence improvement observed in this group.

### 5.2.9. Toxicity Episode Detection

A toxicity episode is operationally defined in this study as a medication change followed by a psychiatric contact within 60 days. This 60-day window was selected to capture clinically proximate events while remaining broad enough to account for delays in care-seeking following treatment changes. This definition captures situations where a change in antipsychotic treatment was temporally associated with clinical deterioration severe enough to require an emergency visit, hospitalization, or outpatient mental health contact. It is important to emphasize that this association is purely temporal and does not imply causality. A psychiatric contact following a medication change may reflect adverse effects of the new treatment, but it may equally reflect the underlying clinical instability that motivated the change in the first place, a relapse unrelated to the medication switch, or routine follow-up scheduled around the transition. The definition therefore captures a signal of clinical vulnerability around medication changes rather than confirmed pharmacological toxicity and should be interpreted accordingly.

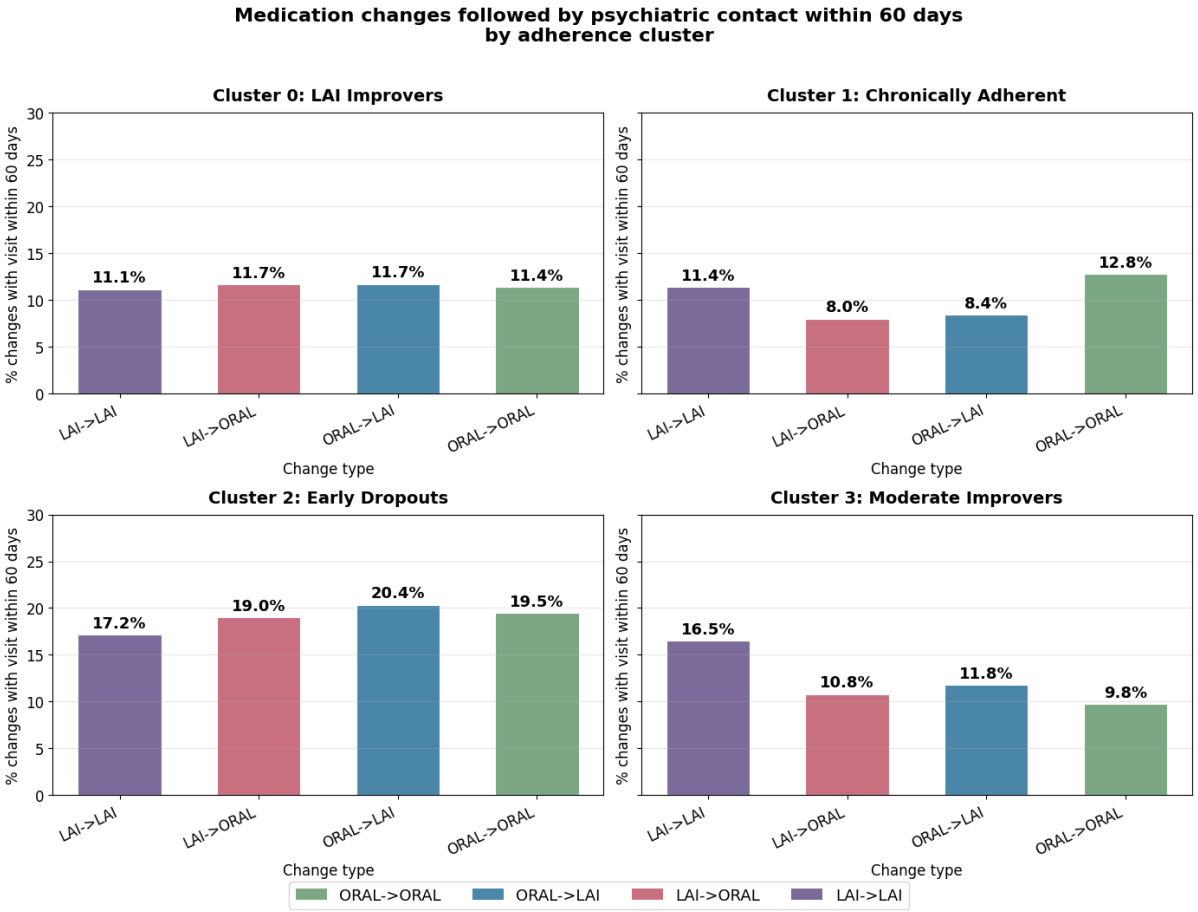



*Figure 17. Percentage of medication changes followed by a psychiatric contact within 60 days out of the total changes of each type, disaggregated by transition type and adherence cluster. Each bar represents the proportion of changes of that type associated with a clinical contact in the subsequent 60-day window across the post-switch observation period (months +1 to +12).*

Medication changes are detected by sorting each patient's dispensing events chronologically and comparing consecutive records. When the dispensed drug's active principle changed between one dispensing and the next, a change event was recorded with the transition type. For each detected change, all three visit tables were searched for any psychiatric contact belonging to the same patient within the 60-day window following the change date. Overall, 11.3% of medication changes were followed by a psychiatric contact within 60 days. The rate was highest in Cluster 2 (Early Dropouts) at 17–20% across all change types, and lowest in Cluster 1 (Chronically Adherent) at 8–13% [Fig. 17].

### 5.2.10. PDC Trajectory by Drug type

An additional exploratory analysis stratifies the mean PDC trajectory by the specific antipsychotic compound dispensed, both for the pre-switch oral phase and the post-switch LAI phase, to examine whether certain medications are associated with distinct adherence patterns that may inform clinical prescribing decisions.

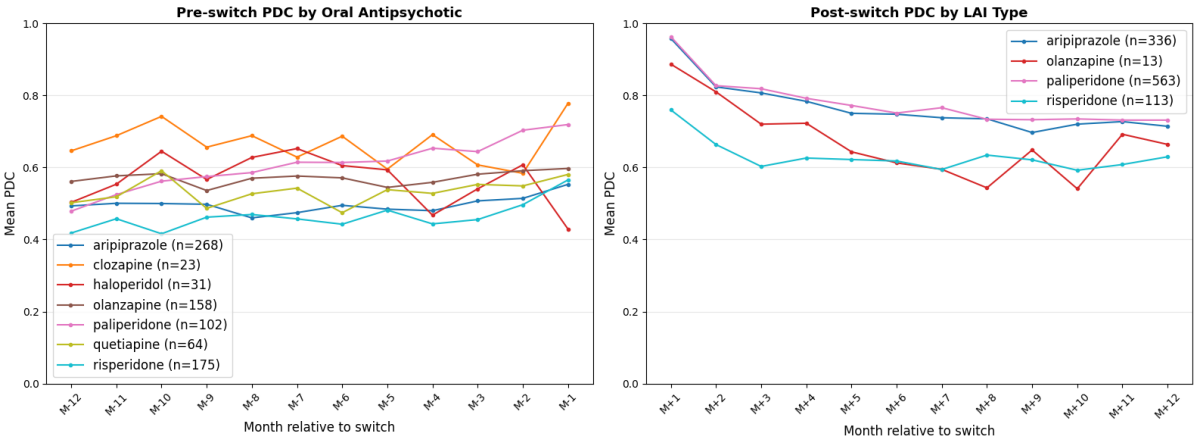



*Figure 18. Mean pre-switch PDC trajectory by main oral antipsychotic dispensed in the 12 months preceding the switch date (left panel) and mean post-switch PDC trajectory by LAI formulation initiated at the switch date (right panel), for the 1,025 patients in the trajectory analysis cohort. Only named active substances are shown in the pre-switch panel; unclassified products are excluded. Sample sizes reflect the number of patients in each drug group. Source: own elaboration.*

In the pre-switch period, 995 patients (97.1%) were primarily treated with second-generation oral antipsychotics, consistent with current prescribing practice in Catalonia, where second-generation agents have largely replaced first-generation compounds as the standard of care for schizophrenia [33]. In the post-switch period, four LAI formulations were represented in the cohort: paliperidone palmitate (n=563, 54.9%), aripiprazole (n=336, 32.8%), risperidone (n=113, 11.0%), and olanzapine pamoate (n=13, 1.3%) [Fig. 18], reflecting the real-world prescribing landscape during the 2015–2017 switch window. Haloperidol decanoate, while classified as an eligible LAI in the study protocol, was not initiated in any patient within the observation window, consistent with its declining use in favor of second-generation LAI formulations [34].

## 5.3. Predictive Modelling: Identifying candidate patients for LAI switching

The predictive modelling component aimed to determine whether the four adherence trajectory clusters identified could be anticipated from pre-switch information alone. A multiclass classification approach was adopted, predicting membership in one of the four trajectory groups rather than applying a binary improver versus non-improver distinction. This choice was deliberate: collapsing the four profiles into two categories would discard clinically meaningful distinctions between patients. Each of these profiles carries different clinical implications and calls for a different management strategy. The model was therefore designed to provide a more granular and actionable risk stratification, applicable at the point of the switch decision using only information available in routine administrative records.

### 5.3.1. Feature Matrix Construction

The feature matrix was assembled by merging multiple PADRIS source tables for around 1,025 patients in the trajectory analysis cohort. The final matrix has shape (1025, 49): 1,025 patients and 49 features after one-hot encoding of categorical variables. No missing values were present across any feature. The 12 monthly PDC values covering months -12 to -1 were taken directly from the PDC pipeline output representing the complete temporal structure of each patient's oral adherence behavior in the year preceding the switch. The switch year was extracted from the first LAI dispensing date and encoded as a categorical variable. Three count features were derived from the psychiatric contact tables: number of primary care visits, psychiatric unit visits, and emergency attendances in the 12 months preceding the

switch. Patients with no recorded contacts were assigned to zero. Sex, age range, socioeconomic level, and smoking status were merged from the sociodemographic table. BMI was excluded due to insufficient coverage (12.2%) across the cohort as it only had data for the year 2020.

The main oral antipsychotic dispensed in the pre-switch period was determined from Table 23 and mapped to active substance categories. The most common were aripiprazole (26.1%), risperidone (17.1%), olanzapine (15.4%), paliperidone (9.9%), and quetiapine (6.2%). Antipsychotic polypharmacy was operationalized as the number of distinct antipsychotic ATC subgroups dispensed in the pre-switch year, ranging from 1 to 6 with a mean of 2.8 (SD 1.1). This reflects antipsychotic polypharmacy specifically rather than general polypharmacy, as the dispensing table was restricted to N05A compounds. LAI type was initially considered a feature but was excluded from the final model, as it is unknown at the time of the switch decision and therefore not applicable in a prospective clinical setting. Its contribution to model performance was confirmed to be negligible (total feature importance <1%).

### 5.3.2. Hyperparameter Optimization and Model Evaluation

The dataset was divided into a training set of 820 patients (80%) and a held-out test set of 205 patients (20%) using stratified sampling to preserve cluster proportions across both sets. Three tree-based ensemble classifiers were evaluated: Random Forest, XGBoost, and LightGBM. This family of models was preferred over alternatives such as logistic regression or neural networks because they handle mixed feature types natively, model non-linear interactions without explicit feature engineering, and produce interpretable feature importance scores [30]. All three models were configured with balanced class weighting to compensate for the moderate size differences between clusters.

Hyperparameter tuning was conducted using Optuna with the Tree-structured Parzen Estimator algorithm, running 20 trials per model evaluated through stratified 5-fold cross-validation on the training set, optimizing macro-averaged one-versus-rest AUC. The parameters optimized were: number of estimators, maximum tree depth, and learning rate for the boosting models. The best cross-validated AUC scores were 0.904 for Random Forest (n_estimators=206, max_depth=11), 0.895 for XGBoost, and 0.894 for LightGBM.

| Model | AUC | Accuracy | Recall (macro) | Precision (macro) |
|---|---|---|---|---|
| Random Forest | 88.70% | 75.12% | 69.92% | 65.27% |
| XGBoost | 88.66% | 73.17% | 66.33% | 58.64% |
| LightGBM | 88.21% | 63.96% | 64.91% | 68.78% |

*Table 8. Test set performance metrics for the three tree-structured candidate models. Yellow highlighted the selected model. Source: self-made.*

Random Forest was selected as the final model with a test AUC of 88.70%, accuracy of 75.12%, macro-averaged recall of 69.92%, and precision of 65.27% [Table 8]. Performance was consistent across all three models, with AUC values within 1.5 percentage points, indicating that the predictive signal in the feature set is robust and not architecture dependent. The moderate recall reflects the inherent difficulty of distinguishing the two intermediate trajectory profiles, Early Dropouts and Moderate Improvers, whose pre-switch PDC patterns overlap more than those of the extreme clusters.

It should be noted that the optimization objective was macro-averaged AUC rather than accuracy, recall, or precision. AUC was chosen because it evaluates the model's discriminative ability across all classification thresholds simultaneously [35], making it more robust to class imbalance than threshold-

dependent metrics. Accuracy, recall, and precision are reported as complementary descriptors of model behavior on the held-out test set, with accuracy being a less informative metric in the presence of unequal cluster sizes.

### 5.3.3. Feature Importance Analysis

Gain-based feature importance was extracted from the Random Forest model, measuring the average improvement in the loss function contributed by each feature across all splits. The 12 pre-switch PDC monthly values collectively accounted for approximately 72% of total gain importance, with months -10 through -11 carrying the highest individual scores [Fig. 19]. This suggests that adherence behavior in the earlier part of the pre-switch year is more discriminative than in the most recent months.

Among non-PDC features, antipsychotic polypharmacy (n_unique_atc_pre) ranked highest (0.024), followed by outpatient mental health visit frequency (0.020), switch year 2016 (0.011), current smoking status (0.010), and risperidone oral type (0.010). Socioeconomic level also appeared among the top predictors.

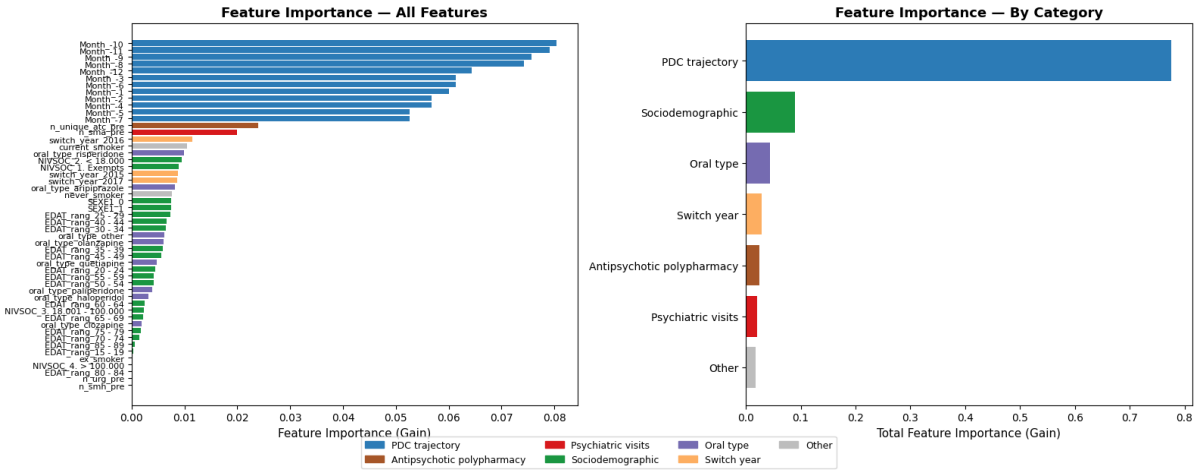



*Figure 19. Gain-based feature importance from the Random Forest classifier , aggregated by feature category (right panel) and shown individually for non-PDC features (left panel). PDC trajectory dominates with approximately 72% of total importance. Among non-PDC features, sociodemographic variables, oral antipsychotic type, and switch year contribute the most, while psychiatric visits and antipsychotic polypharmacy show modest contributions. Source: own elaboration.*

### 5.3.4. SHAP Analysis by Cluster

SHAP (SHapley Additive exPlanations) values were computed to characterize the directional influence of non-PDC features on cluster assignment. Unlike feature importance, which provides a single aggregate score per feature, SHAP values are computed per patient per feature, allowing the direction and heterogeneity of each feature's influence to be visualised. A Random Forest classifier was trained exclusively on the non-PDC features to isolate the contribution of clinical and sociodemographic variables independently of the dominant PDC signal, addressing the clinically relevant question of what can be predicted from patient characteristics alone. Beeswarm plots were generated [Fig. 20] for each of the four clusters, displaying the top 15 non-PDC features ranked by mean absolute SHAP value. Each point represents one patient, with the x-axis showing the SHAP value, positive values push the prediction towards that cluster and color encoding the feature value from low (blue) to high (red).

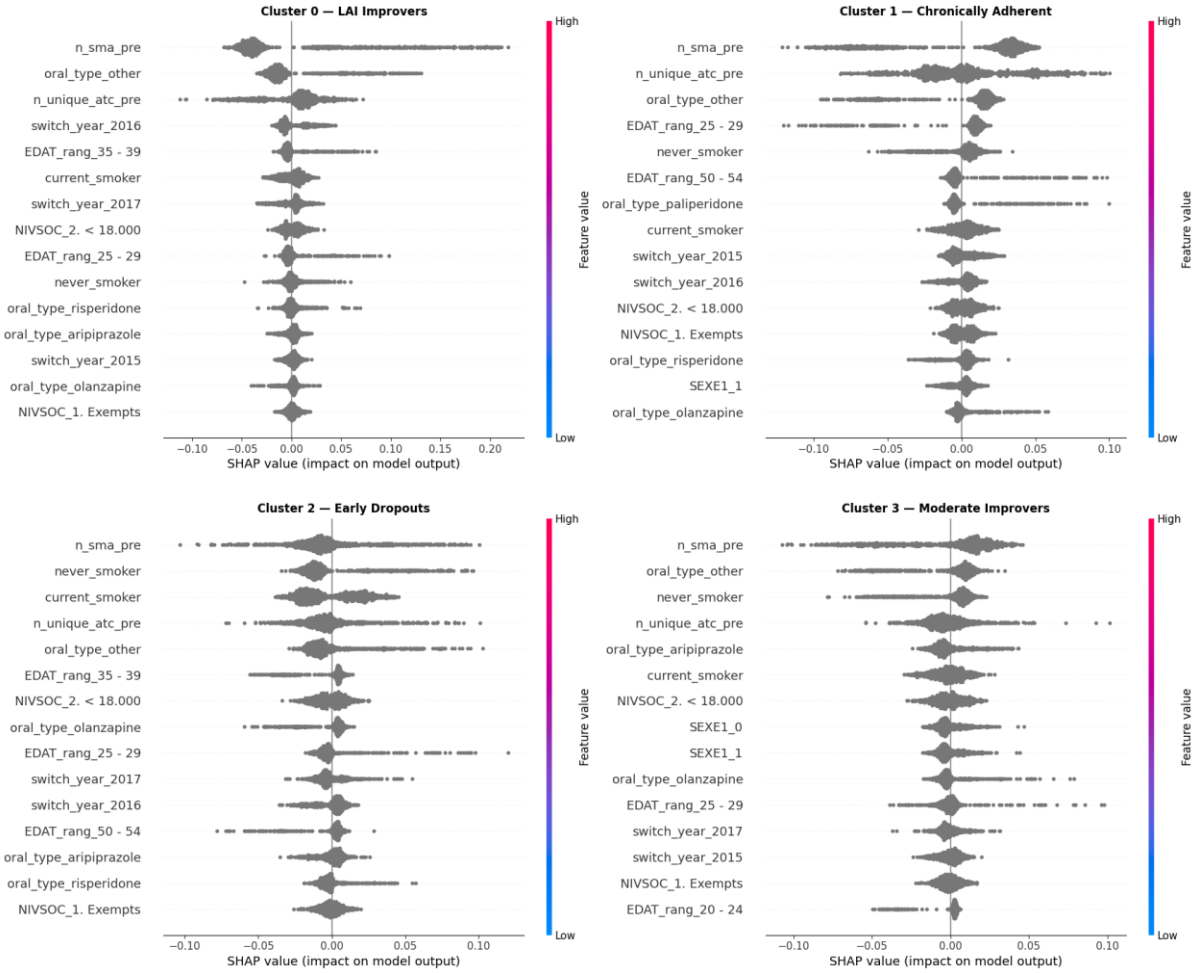



*Figure 20. SHAP beeswarm plots for each of the four trajectory clusters, computed from a Random Forest classifier trained on the 38 non-PDC features. Each point represents one patient. The x-axis shows the SHAP value, where positive values indicate that the feature pushes the prediction towards that cluster. Point colour encodes the feature value from low (blue) to high (red). Features are ranked by mean absolute SHAP value. Source: own elaboration.*

For Cluster 0 (LAI Improvers), higher outpatient visit frequency (n_sma_pre) and other oral types carry positive SHAP values. For Cluster 1 (Chronically Adherent), higher antipsychotic polypharmacy, paliperidone and age range 50-54 positively associated, with never-smoker showing a negative contribution. For Cluster 2 (Early Dropouts), higher outpatient visits and never smokers show positive associations. For Cluster 3 (Moderate Improvers), SHAP values are generally smaller in magnitude, consistent with the intermediate nature of this trajectory group.

### 5.3.5. Prospective Application to Oral-Only Patients

The trained Random Forest classifier was applied to the 5,662 patients in the PADRIS cohort who received antipsychotic dispensing throughout the observation period but never initiated LAI treatment. For these patients, the reference date was defined as each patient's last recorded dispensing date, and the PDC was computed over the 12 preceding months using the same monthly calculation logic as for the trajectory cohort creating a matrix with dimensions (5662, 13). The switch year was imputed as 2016, the most frequent value in the training cohort, given its modest contribution to model performance.

The resulting PDC distribution showed a mean of 0.61 (SD 0.40) and a median of 0.80 (IQR 0.17-1.00), with 858 patients (15.2%) recording a mean PDC of zero, indicating complete non-engagement with pharmacy dispensing in their final observation year. The model predicted 1,455 patients (25.7%) as

Cluster 0 (LAI Improvers), somewhat below the 31.3% observed in the training cohort, suggesting the model is conservative in its identification of improver candidates in the non-switcher population. The remaining patients were distributed across Cluster 1 (Chronically Adherent, n=3,094, 54.6%), Cluster 2 (Early Dropouts, n=737, 13.0%), and Cluster 3 (Moderate Improvers, n=376, 6.6%). The 1,455 identified patients represent a prioritized subgroup for whom a clinical review of LAI initiation may be warranted, providing a data-driven complement to routine clinical assessment.

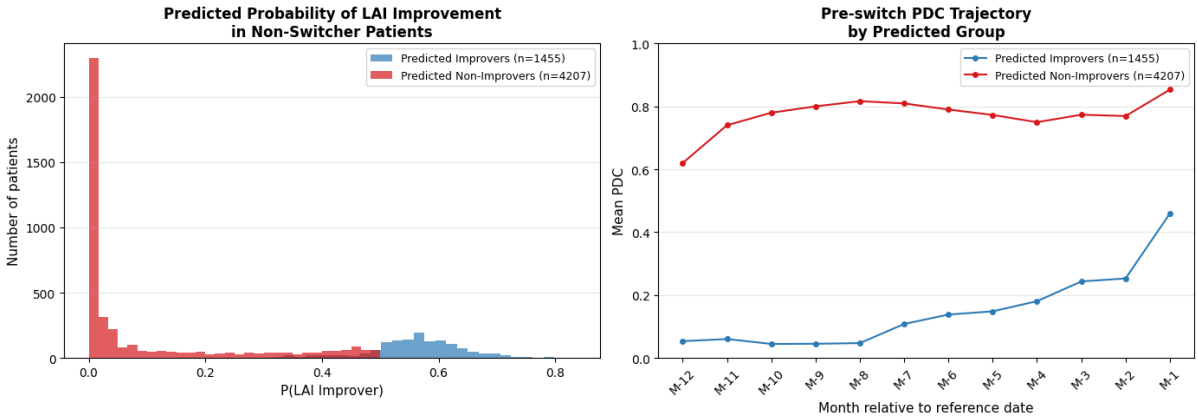



*Figure 21. Predicted probability of LAI improvement (left) and mean pre-switch PDC trajectory (right) for the 5,662 oral-only non-switcher patients, stratified by predicted group (LAI Improvers vs Non-Improvers). Source: own elaboration.*

Predicted non-improvers are concentrated near P=0, reflecting high model confidence, while predicted improvers spread between 0.5 and 0.8 [Figure 21]. The PDC trajectories confirm clinical coherence that the predicted improvers show consistently low oral adherence throughout the reference window, while predicted non-improvers maintain high adherence.

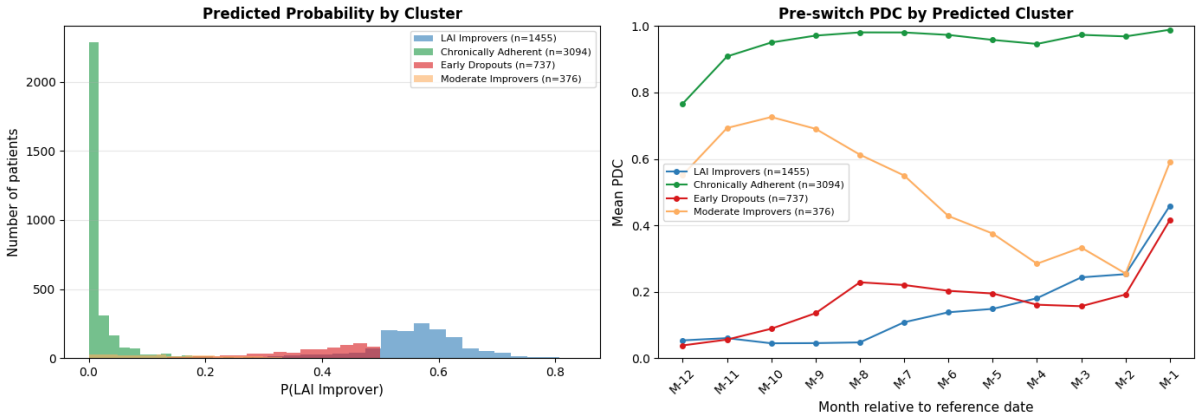



*Figure 22. Predicted probability of LAI improvement (left) and mean pre-switch PDC trajectory (right) for the 5,662 oral-only non-switcher patients, stratified by all four predicted trajectory clusters. Source: own elaboration.*

The PDC trajectories show clear separation between clusters [Fig. 22], with Chronically Adherent patients maintaining near-perfect adherence throughout and LAI Improvers showing persistently low adherence with a sharp rise in the final month. Early Dropouts and Moderate Improvers show intermediate patterns.

Clinical coherence was validated by inspecting the mean PDC trajectories by predicted cluster. Patients predicted as LAI Improvers showed consistently low oral adherence throughout the reference window, mirroring the training cohort profile, while patients predicted as Chronically Adherent maintained high oral PDC throughout. The probability distribution of P(LAI Improver) showed a bimodal structure, with

predicted non-improvers concentrated near zero and predicted improvers spread between 0.5 and 0.8, consistent with the inherent uncertainty of a prospective prediction task.

The 1,895 identified patients represent a prioritized subgroup for whom a clinical review of LAI initiation may be warranted, providing a data-driven complement to routine clinical assessment.

# 6. Execution Schedule

## 6.1 Work Breakdown Structure (WBS)

The success of a project relies significantly on effective planning and detailed execution. One way to achieve this is by subdividing the work into smaller, more manageable components. This has been accomplished using the Work Breakdown Structure (WBS), a key project management tool that organizes all tasks within the project's scope. In this project, the WBS categorizes the work into six main packages: project administration, literature review, data management, pipeline development and implementation, data analysis, and project closure.

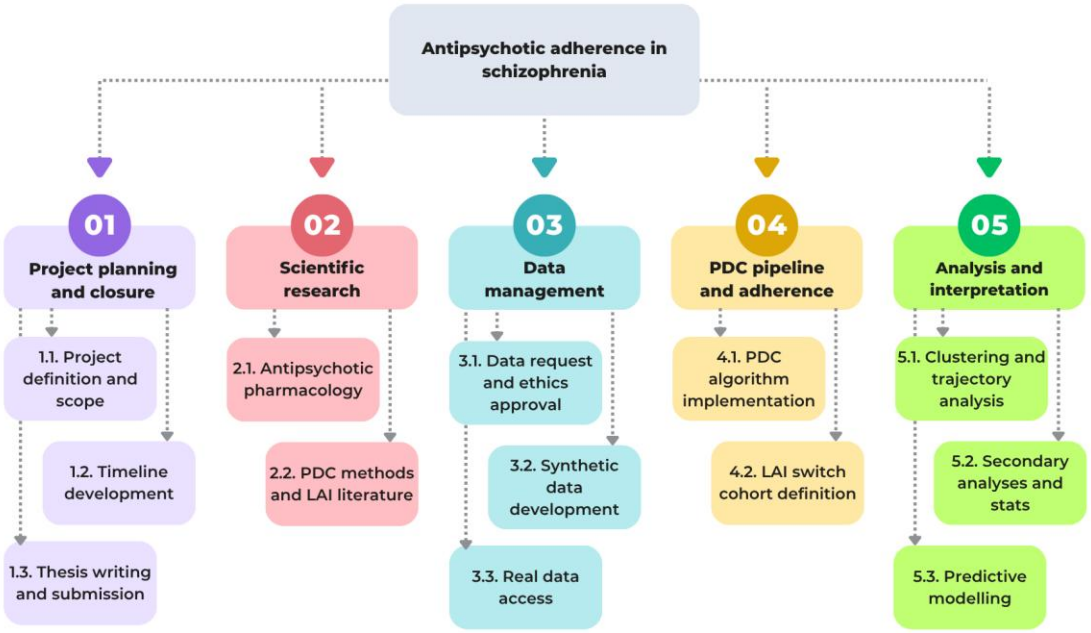



*Figure 23. WBS diagram for the project, illustrating the main work packages and their associated tasks. Source: own elaboration.*

This structured approach not only clarifies the sequence of tasks but also facilitates effective monitoring of the project's progress. Figure 23 presents the WBS developed for this study, offering an overview of the project's work packages and their associated tasks.

## 6.2 Execution Chronogram (GANTT)

In addition to the WBS, a temporal chronogram was developed to assess the timeline of tasks and their interdependencies. This tool is key to visually interpreting the project sequence, identifying activities that are critical and must adhere to strict timelines, and distinguishing those that allow for some flexibility.

The project spans from December 2025 to June 2026. The initial phase covers administrative tasks and data request approval, followed by a literature review conducted in parallel across the first two months. Synthetic data development begins in February 2026 to enable early pipeline testing prior to the arrival of real data, which is marked as a key milestone in March 2026. Pipeline development and adherence algorithm implementation are then carried out, followed by clustering and trajectory analysis, secondary

analyses, and predictive modelling, all concentrated in the final months. Thesis writing runs concurrently with the analytical phases, and submission is set as the final milestone at the end of June 2026.

| ID | Activity | Precedents | Duration (hours) |
|---|---|---|---|
| A | 1.1. Project Definition and Scope | - | 10 |
| B | 1.2. Timeline development | A | 5 |
| C | 2.1. Antipsychotic pharmacology research | A | 20 |
| D | 2.2. PDC methods and LAI literature | A | 25 |
| E | 3.1. Data request and ethics approval | B | 5 |
| F | 3.2.Synthetic data development | C, D | 50 |
| G | 3.3. Real data access | E | 5 |
| H | 4.1. LAI switch cohort definition | G | 20 |
| I | 4.2. PDC algorithm implementation | H | 60 |
| J | 5.1. Clustering and trajectory analysis | I | 60 |
| K | 5.2. Secondary analyses and stats | J | 35 |
| L | 5.3. Predictive modelling | K | 15 |
| M | 1.3. Thesis writing and submission | L | 80 |
| N | 1.4. Thesis presentation | M | 10 |

*Table 9. Activities with their duration and precedents in hours. Total: 400 hours.*

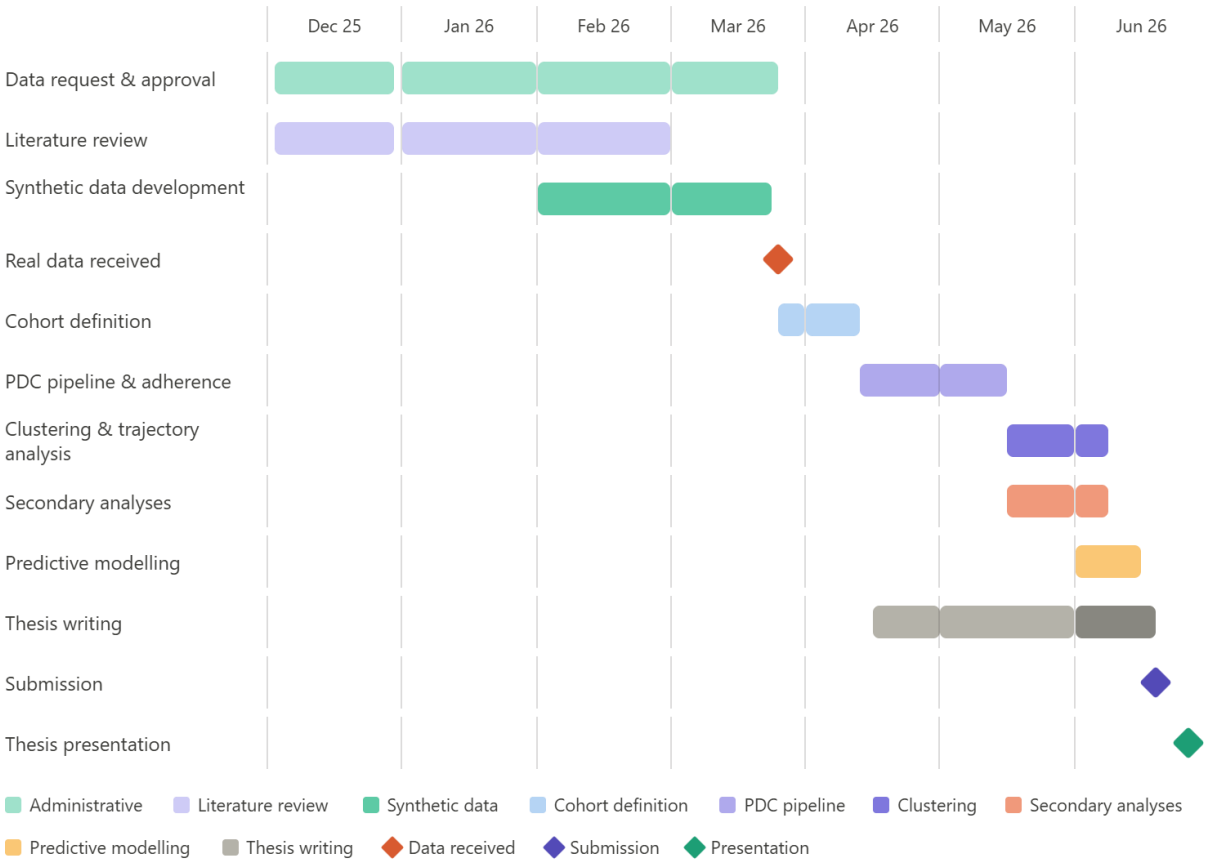



*Figure 24. Clear visualization of the overall plan, aiding in task management and monitoring throughout the project. Source: own elaboration*

## 6.3 Critical Path Method Diagram (PERT)

Complementary to the GANTT diagram, the PERT or Critical Path Method diagram provides a visual representation of the chronology and interdependencies between activities. Activities are represented as arrows connecting numbered nodes, each displaying the earliest and latest times at which that stage can begin without delaying the overall project. The critical path, comprising activities with zero slack that must be completed on time, is highlighted in red, while flexible activities that can tolerate delays without affecting the final deadline are displayed in black.

Figure 25 illustrates the PERT diagram developed for this project.

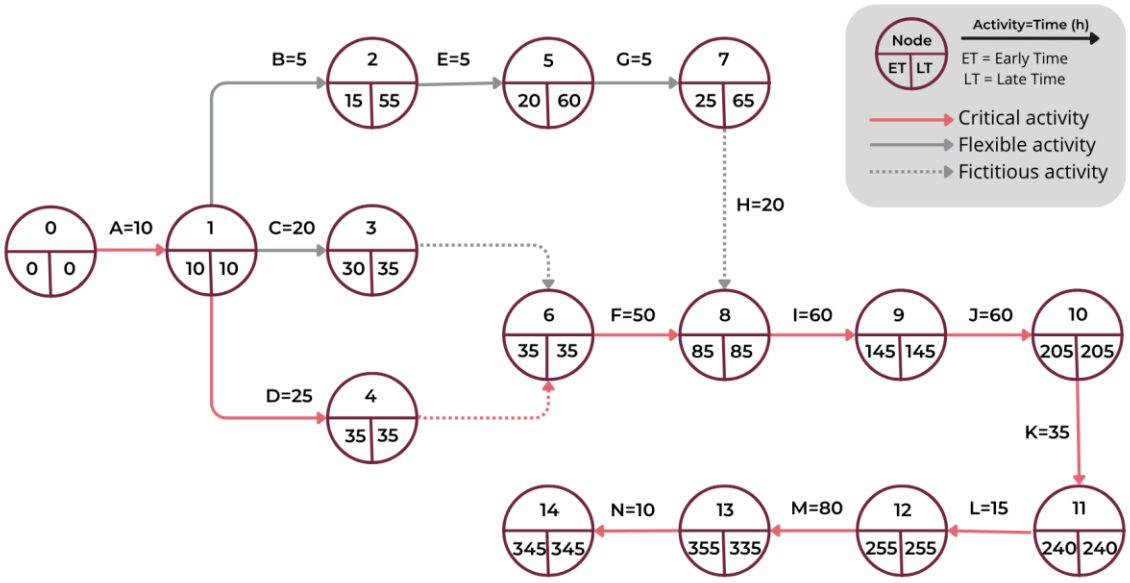



*Figure 25. PERT diagram illustrating the chronology and interdependencies of activities within the project. Source: own elaboration*

# 7. Technical Viability

This section evaluates the technical viability of the analytical pipeline developed in this project, identifying factors that have influenced its development positively and negatively, as well as those that may affect its future application and scalability. This analysis is presented in a SWOT matrix (Strengths, Weaknesses, Opportunities, Threats) summarized in Table 9, which outlines the key dimensions of the project from a technical and translational perspective.

Strengths include population-scale data coverage through the PADRIS system, which minimizes selection bias and enables statistically robust analyses, alongside a modular and reproducible Python pipeline that can be adapted to other psychiatric conditions or extended to new data releases with limited additional development. Weaknesses include the administrative and legal complexity of accessing PADRIS data, the inherent limitations of dispensing-based adherence metrics which cannot capture actual medication intake, and the absence of clinical variables such as symptom severity scores or structured medication records that would enrich the predictive models. Opportunities lie in the potential to standardize this pipeline for routine pharmacoepidemiological monitoring within the Catalan public health system, and to extend the trajectory-based approach to other chronic psychiatric conditions such as bipolar disorder or treatment-resistant depression. Threats include changes in regional prescription protocols or data governance frameworks that could affect reproducibility, the risk of model degradation over time as prescribing patterns evolve, and the inherent limitations of a proof-of-concept predictive model that has not yet been externally validated.

| | Positive | Negative |
|---|---|---|
| Internal | **Strengths**<br>– Population-scale coverage (1,421,510 individuals) minimising selection bias<br>– Real-world dispensing data reflecting actual clinical practice in Catalonia<br>– Modular and reproducible Python pipeline adaptable to other conditions or data releases<br>– Longitudinal follow-up of up to 10 years (2010–2019)<br>– Collaboration with IDIBAPS and Hospital Clínic de Barcelona providing clinical expertise<br>– Unsupervised clustering approach that does not impose prior assumptions on adherence patterns | **Weaknesses**<br>– Dispensing records do not confirm actual medication intake<br>– Administrative approval process delays access to real data<br>– BMI, smoking status and comorbidities coverage insufficient for full inclusion in predictive models<br>– Predictive model trained and evaluated on the same cohort, no external validation<br>– LAI type required imputation for prospective application<br>– Absence of clinical severity scores or structured psychiatric outcome measures<br>– PDC trajectory data was used both as the clustering input and as the primary outcome measure for characterizing the resulting groups |
| External | **Opportunities** | **Threats** |

| | | |
|---|---|---|
| | – Standardisation of the pipeline for routine pharmacoepidemiological monitoring in Catalonia<br>– Extension to other chronic psychiatric conditions (bipolar disorder, treatment-resistant depression)<br>– Integration with clinical outcome data to formally test adherence-outcome associations<br>– Potential use as a decision-support tool in clinical practice for LAI candidate identification<br>– Growing institutional interest in real-world evidence for treatment evaluation | – Changes in regional prescription protocols may affect generalizability over time<br>– Model degradation as prescribing patterns evolve beyond the 2015–2017 training window<br>– Data governance framework changes affecting future PADRIS access<br>– Risk of misinterpretation of temporal associations as causal relationships<br>– Dependency on completeness of dispensing records which may vary across healthcare regions |

*Table 10. SWOT analysis of the project, with Strengths, Weaknesses, Opportunities and Threats. Source: own elaboration.*

# 8. Economic Viability

This section provides a detailed estimate of the budget required to cover all expenses associated with the complete execution of the project. These costs are specified in Table 11.

| Item | Description | Unit Price | Total Price |
|---|---|---|---|
| Data Acquisition | | | |
| PADRIS-PRESTO data | Accessed under institutional approval | Free | 0€ |
| Software | | | |
| Python | Programming language for data analysis and libraries | Free | 0€ |
| GitHub | Code repository | Free | 0€ |
| Visual Studio Code | Code editor | Free | 0€ |
| Material Resources | | | |
| Personal Laptop | Lenovo laptop, 800€ purchase price, 6 months depreciation over 3-year lifecycle | 133€ | 133€ |
| Institutional computing | Access to computing infrastructure provided by IDIBAPS/Hospital Clínic de Barcelona | Institutional | 0€ |
| Human Resources | | | |
| Biomedical Engineer | Total time: 400 hours | 20€/hour | 8,000€ |
| Tutor | Total time 40 hours | 50€/hour | 2,000€ |
| Total Cost | | | 10,133€ |

*Table 11. Breakdown of costs and resources for the project, including software tools, material resources, and human resources.*

As shown in Table 10, the project costs are divided into four categories: data acquisition, software, material resources, and human resources. The data used in this study was obtained at no direct cost, as access to the PADRIS-PRESTO cohort was granted through the institutional research group at IDIBAPS/Hospital Clínic de Barcelona under an approved research protocol. Any administrative costs associated with the PADRIS data access approval were covered by the research group and are therefore not included here.

All software tools used throughout the project are open-source and freely accessible, resulting in no associated licensing costs. Regarding material resources, only the depreciation of the personal laptop used during the project period is included, calculated as six months of a standard three-year depreciation cycle. Access to additional computing infrastructure at IDIBAPS/Hospital Clínic de Barcelona was provided as part of the institutional collaboration at no cost to the project.

The largest cost component corresponds to human resources, comprising 400 hours of work by the biomedical engineer at a reference rate of 20€/hour, and 40 hours of supervision by the project tutor at 50€/hour, for a combined total of 10,000€. The overall estimated project cost is 10,133€.

# 9. Regulations and Legal Aspects

The study protocol was submitted to the Drug Research Ethics Committee (Comité d'Ètica de la Investigació amb Medicaments, CEIm) of Hospital Clínic de Barcelona and received formal approval on 11 March 2026. Given the nature of the study, no specific data collection form was required; the project relies exclusively on the secondary use of pseudonymized data from the PADRIS platform, as communicated to the committee prior to approval.

The study was conducted in accordance with the Declaration of Helsinki (Fortaleza, 2024 revision) and with Royal Decree 957/2020, which regulates observational studies with medicinal products for human use in Spain. All data handling complied with Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation, GDPR) and with Spanish Organic Law 3/2018 on Personal Data Protection and Digital Rights Guarantee (LOPDGDD). The data used in this study is pseudonymized and originates from the PADRIS-PRESTO cohort; no information enabling patient re-identification was included at any stage of the analysis.

Data transfer from the PRESTO research team to the analytical team was carried out exclusively through the secure institutional servers of Hospital Clínic de Barcelona, ensuring full traceability and restricting access to the investigators named in the approved protocol. All processing was performed within the institution's secure computing environment, with no transfer of data to third parties.

In accordance with regulatory requirements, the investigator and sponsor are committed to retaining the study data for a minimum of five years following its completion.

# 10. Conclusions and Future Lines

This thesis developed and applied a population-based data analysis pipeline to characterize antipsychotic adherence trajectories in patients with schizophrenia in Catalonia, using real-world dispensing data from the PADRIS-PRESTO cohort. A well-defined analytical cohort of 1,025 patients with a confirmed schizophrenia diagnosis who switched from oral to LAI antipsychotic treatment between 2015 and 2017 was successfully identified through a six-step sequential selection funnel.

Monthly PDC computation across the full 24-month observation window revealed a population-level adherence improvement following LAI initiation, with the mean post-switch PDC rising from approximately 50% to a peak of 93% in the first post-switch month before stabilising around 70%. The k=4 TimeSeriesKMeans clustering solution identified four clinically distinct trajectory profiles: LAI Improvers (31.3%), who showed dramatic adherence gains from a very low pre-switch baseline; Chronically Adherent patients (31.1%), who maintained high adherence throughout; Early Dropouts (18.2%), who initiated LAI but discontinued within months; and Moderate Improvers (19.3%), who achieved borderline adherence around the 80% threshold. These findings confirm that summarising adherence as a single aggregate metric obscures clinically meaningful variation in how patients respond to LAI switching.

Psychiatric contact patterns differed markedly across clusters. LAI Improvers showed the largest reduction in contacts across all service types following the switch, consistent with clinical stabilisation. Early Dropouts maintained persistently elevated inpatient contacts post-switch, reflecting ongoing clinical instability. As a proof-of-concept component, a multiclass Random Forest classifier achieved a test AUC of 88.7% in predicting trajectory cluster membership from pre-switch information alone and identified 1,455 oral-only patients (25.7%) as likely LAI Improver candidates when applied prospectively.

These results provide population-level evidence that LAI switching is associated with adherence improvement in a substantial proportion of patients with schizophrenia in Catalonia, but that this benefit is heterogeneous and not universal. Patients who were already adherent to oral treatment did not show meaningful adherence gains following the switch, while those with chronically poor oral adherence benefited most. The Early Dropout profile highlights that LAI initiation alone is insufficient to guarantee sustained adherence in clinically vulnerable patients, and that targeted retention interventions may be warranted for this subgroup. The trajectory-based framework developed here offers a methodologically rigorous and clinically interpretable approach to characterizing this heterogeneity from routine administrative data.

This study has several limitations. First, adherence was estimated using dispensing records and therefore reflects medication availability rather than confirmed intake, a patient who collected their medication but did not take it would be classified as adherent. Second, selection bias may be present, as patients switching to LAI may differ systematically from those remaining on oral therapy in ways not fully captured by the available covariates. Third, administrative data lack detailed clinical variables such as symptom severity scores, patient-reported outcomes, or reasons for treatment change, which limits the interpretability of adherence differences across clusters. Fourth, the predictive model was trained and evaluated on the same cohort from which the cluster labels were derived, meaning external validation on an independent population is required before any clinical deployment. Finally, the restriction of the

observation window to 2015–2017 ensures data completeness but limits generalizability to more recent prescribing patterns, particularly following the broader adoption of newer LAI formulations.

The pipeline developed in this thesis is modular, reproducible, and designed to be extensible to other psychiatric conditions and future PADRIS data releases. It demonstrates that population-level dispensing data, when analyzed through a trajectory-based framework, can provide actionable clinical insights into the heterogeneous real-world effectiveness of LAI antipsychotics, insights that aggregate metrics alone cannot deliver.

**Future Lines:**

- External validation of the predictive model on an independent cohort of switchers from a different time period or healthcare region
- Extension of the trajectory clustering approach to mental health conditions such as bipolar disorder or treatment-resistant depression within the PADRIS system
- Linkage of adherence trajectories to hard clinical outcomes including mortality, hospitalization rates, and functional recovery measures, to formally test whether trajectory group membership predicts long-term prognosis
- Integration of clinical variables such as symptom severity scores, comorbidity profiles and complete BMI variation data to enrich the predictive feature set beyond administrative data
- Development of a real-time decision-support tool that stratifies LAI candidates at the point of clinical decision-making using routinely available dispensing and sociodemographic data
- Investigation of the Early Dropout profile in greater depth, including identification of modifiable risk factors and evaluation of targeted retention interventions

# 11. Bibliography

[1] Freedman, R. (2003). Schizophrenia. New England Journal of Medicine, 349(18), 1738–1749. https://doi.org/10.1056/nejmra035458

[2] Haddad, P., Brain, C., & Scott, J. (2014). Nonadherence with antipsychotic medication in schizophrenia: challenges and management strategies. Patient Related Outcome Measures, 5(5), 43. https://doi.org/10.2147/prom.s42735

[3] Kane, J. M., Ofer Agid, Castle, D. J., Citrome, L., Fagiolini, A., Kishimoto, T., Larrauri, C. A., Leucht, S., Rubio, J. M., Sajatovic, M., Schooler, N., & Correll, C. U. (2025). The Use of Long-Acting Injectables for People with Schizophrenia: Consensus Panel Recommendations for Overcoming Barriers and Implementing Treatment. Neurology and Therapy. https://doi.org/10.1007/s40120-025-00838-3

[4] Boyer, L., Falissard, B., Nuss, P., Collin, C., Duret, S., Rabbani, M., De Chefdebien, I., Tonelli, I., Llorca, P. M., & Fond, G. (2023). Real-world effectiveness of long-acting injectable antipsychotic treatments in a nationwide cohort of 12,373 patients with schizophrenia-spectrum disorders. Molecular Psychiatry, 1–8. https://doi.org/10.1038/s41380-023-02175-z

[5] *Notes de premsa*. (2022). Canal Salut. https://canalsalut.gencat.cat/ca/inici/nota-premsa/?id=445862

[6] De Prisco, M., Oliva, V., Fico, G., Mas, A., Valenzuela-Pascual, C., Montejo, L., Bort, M., Sommerhoff, C., Bortolozzi, A., Miquel-Rio, L., Vilella, E., Forte, M. F., Fortea, L., Fernández, T., Giménez-Palomo, A., Sague Vilavella, M., Madero, S., Llorca Bofí, V., Bioque, M., & Grande, I. (2025). The PADRIS-PRESTO cohort: A comprehensive population-based study on mental health in Catalonia. *European Psychiatry : The Journal of the Association of European Psychiatrists*, *68*(1), e144. https://doi.org/10.1192/j.eurpsy.2025.10103

[7] SCHULTZ, S. H., NORTH, S. W., & SHIELDS, C. G. (2007). Schizophrenia: A Review. American Family Physician, 75(12), 1821–1829. https://www.aafp.org/afp/2007/0615/p1821

[8] Häfner, H. (2019). From onset and prodromal stage to a life-long course of schizophrenia and its symptom dimensions: How sex, age, and other risk factors influence incidence and course of illness. Psychiatry Journal, 2019(1), 1–15. https://doi.org/10.1155/2019/9804836

[9] Kendler, Kenneth S. “Phenomenology of Schizophrenia and the Representativeness of Modern Diagnostic Criteria.” JAMA Psychiatry, vol. 73, no. 10, 1 Oct. 2016, p. 1082, https://doi.org/10.1001/jamapsychiatry.2016.1976

[10] Mas, A., Clougher, D., Anmella, G., Valenzuela-Pascual, C., De Prisco, M., Oliva, V., Fico, G., Grande, I., Morilla, I., Segú, X., Primé-Tous, M., Ruíz, V., Also, M. A., Murgui, S., Sant, E., Sans-Corrales, M., Fullana, M. À., Sisó-Almirall, A., Radua, J., & Blanch, J. (2024). Trends and associated factors of mental health diagnoses in Catalan Primary Care (2010-2019). European Psychiatry : The Journal of the Association of European Psychiatrists, 67(1), e81. https://doi.org/10.1192/j.eurpsy.2024.1793

[11] Lewis, D. A., & Lieberman, J. A. (2000). Catching Up on Schizophrenia. *Neuron*, *28*(2), 325–334. https://doi.org/10.1016/s0896-6273(00)00111-2

[12] Bzdok, Danilo, and Andreas Meyer-Lindenberg. "Machine Learning for Precision Psychiatry: Opportunities and Challenges." Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, vol. 3, no. 3, Mar. 2018, pp. 223–230, www.biologicalpsychiatrycnni.org/article/S2451-9022(17)30206-9/fulltext, https://doi.org/10.1016/j.bpsc.2017.11.007

[13] Tienari, P., Wynne, L. C., Sorri, A., Lahti, I., Läksy, K., Moring, J., Naarala, M., Nieminen, P., & Wahlberg, K.-E. (2004). Genotype–environment interaction in schizophrenia-spectrum disorder. *British Journal of Psychiatry*, *184*(3), 216–222. https://doi.org/10.1192/bjp.184.3.216

[14] Trubetskoy, Vassily, et al. "Mapping Genomic Loci Implicates Genes and Synaptic Biology in Schizophrenia." Nature, vol. 604, no. 7906, 1 Apr. 2022, pp. 502–508, www.nature.com/articles/s41586-022-04434-5, https://doi.org/10.1038/s41586-022-04434-5

[15] Rodriguez, Victoria, et al. "Polygenic and Polyenvironment Interplay in Schizophrenia-Spectrum Disorder and Affective Psychosis; the EUGEI First Episode Study." Schizophrenia Bulletin, 10 Dec. 2024, academic.oup.com/schizophreniabulletin/advance-article/doi/10.1093/schbul/sbae207/7920767, https://doi.org/10.1093/schbul/sbae207. Accessed 20 Dec. 2024.

[16] Mas, Sergi, et al. "Examining Gene–Environment Interactions Using Aggregate Scores in a First-Episode Psychosis Cohort." Schizophrenia Bulletin, vol. 46, no. 4, 21 Feb. 2020, pp. 1019–1025, https://doi.org/10.1093/schbul/sbaa012. Accessed 28 Nov. 2023.

[17] Avasthi, A., Grover, S., Chakrabarti, S., & Kulhara, P. (2017). Clinical Practice Guidelines for Management of Schizophrenia. Indian Journal of Psychiatry, 59(5), 19. https://doi.org/10.4103/0019-5545.196972

[18] Leucht, Stefan, et al. "Schizophrenia." Nature Reviews Disease Primers, vol. 11, no. 1, 27 Nov. 2025, www.nature.com/articles/s41572-025-00667-6?utm_source, https://doi.org/10.1038/s41572-025-00667-6.

[19] A systematic review of the indirect costs of schizophrenia in Europe. (2024). Nih.gov. https://pmc.ncbi.nlm.nih.gov/articles/PMC6241204/

[20] Oliva, J., López-Bastida, J., Rubén Osuna-Guerrero, Angel Luis Montejo-González, & B. Duque-González. (2006). The costs of schizophrenia in Spain. *European Journal of Health Economics*, *7*(3), 179–184. https://doi.org/10.1007/s10198-006-0350-5

[21] Anmella, G., Sanabra, M., Primé-Tous, M., Segú, X., Solanes, A., Ruíz, V., Morilla, I., Also Fontanet, A., Sant, E., Murgui, S., Sans-Corrales, M., Martínez-Aran, A., Fico, G., De Prisco, M., Oliva, V., Murru, A., Zahn, R., Young, A. H., Vicens, V., & Viñas-Bardolet, C. (2022). Antidepressants overuse in primary care: Prescription trends between 2010 and 2019 in Catalonia. Revista de Psiquiatría Y Salud Mental. https://doi.org/10.1016/j.rpsm.2022.12.00121]

[22] Kotzeva, A., Mittal, D., Desai, S., Judge, D., & Samanta, K. (2022). Socioeconomic burden of schizophrenia: a targeted literature review of types of costs and associated drivers across 10

countries. *Journal of Medical Economics*, *26*(1), 70–83. https://doi.org/10.1080/13696998.2022.2157596

[23] Leire Urien, Unax Lertxundi, Garcia, M., Aguirre, C., Nerea Jauregizar, & Morera-Herreras, T. (2025). Oral Adverse Effects of Antipsychotic Medications: A Case/Noncase Analysis of EudraVigilance Data. *Oral Diseases*. https://doi.org/10.1111/odi.15337

[24] Clínica Universidad de Navarra (2023). *Discinesia*. https://www.cun.es/diccionario-medico/terminos/discinesia

[25] Zhang, J.-P., Gallego, J. A., Robinson, D. G., Malhotra, A. K., Kane, J. M., & Correll, C. U. (2013). Efficacy and safety of individual second-generation vs. first-generation antipsychotics in first-episode psychosis: a systematic review and meta-analysis. *International Journal of Neuropsychopharmacology*, *16*(6), 1205–1218. https://doi.org/10.1017/s1461145712001277

[26] Hasan, A., Falkai, P., Wobrock, T., Lieberman, J., Glenthoj, B., Gattaz, W. F., Thibaut, F., & Möller, H.-J. (2012). World Federation of Societies of Biological Psychiatry (WFSBP) Guidelines for Biological Treatment of Schizophrenia, Part 2: Update 2012 on the long-term treatment of schizophrenia and management of antipsychotic-induced side effects. *The World Journal of Biological Psychiatry*, *14*(1), 2–44. https://doi.org/10.3109/15622975.2012.739708

[27] Haddad, P. M., & Correll, C. U. (2023). Long-acting antipsychotics in the treatment of schizophrenia: opportunities and challenges. *Expert Opinion on Pharmacotherapy*, *24*(4), 473–493. https://doi.org/10.1080/14656566.2023.2181073

[28] Alhazami, M., Pontinha, V. M., Patterson, J. A., & Holdford, D. A. (2020). Medication Adherence Trajectories: A Systematic Literature Review. *Journal of Managed Care & Specialty Pharmacy*, *26*(9), 1138–1152. https://doi.org/10.18553/jmcp.2020.26.9.1138

[29] MacEwan, J. P., Forma, F., Shafrin, J., Hatch, A., Lakdawalla, D. N., & Lindenmayer, J.-P. (2016). Patterns of Adherence to Oral Atypical Antipsychotics Among Patients Diagnosed with Schizophrenia. *Journal of Managed Care & Specialty Pharmacy*, *22*(11), 1349–1361. https://doi.org/10.18553/jmcp.2016.22.11.1349

[30] Lundberg, Scott M., et al. "From Local Explanations to Global Understanding with Explainable AI for Trees." Nature Machine Intelligence, vol. 2, no. 1, Jan. 2020, pp. 56–67, https://doi.org/10.1038/s42256-019-0138-9

[31] Akiba, Takuya, et al. "Optuna: A Next-Generation Hyperparameter Optimization Framework." Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 25 July 2019, https://doi.org/10.1145/3292500.3330701

[32] European Commission Approves TREVICTA® (paliperidone palmitate a 3-monthly injection), for Maintenance Treatment of Schizophrenia. (2023). Jnj.com. https://www.investor.jnj.com/investor-news/news-details/2016/European-Commission-Approves-TREVICTA-paliperidone-palmitate-a-3-monthly-injection-for-Maintenance-Treatment-of-Schizophrenia-05-31-2016/default.aspx

[33] Gaviria, Ana M., et al. “A Non-Interventional Naturalistic Study of the Prescription Patterns of Antipsychotics in Patients with Schizophrenia from the Spanish Province of Tarragona.” Nature Machine Intelligence, 2024, https://pmc.ncbi.nlm.nih.gov/articles/PMC4591292/

[34] Long-acting injectable antipsychotics for the treatment of schizophrenia in Spain. Revista de Psiquiatría y Salud Mental. 2019. doi:10.1016/j.rpsm.2019.01.002 https://www.sciencedirect.com/science/article/pii/S2173505019300160

[35] Fawcett, Tom. “An Introduction to ROC Analysis.” Pattern Recognition Letters, vol. 27, no. 8, June 2006, pp. 861–874, https://doi.org/10.1016/j.patrec.2005.10.010

## Pyhton Libraries

Data manipulation & simulation
- pandas
- numpy

Time series & dimensionality reduction
- tslearn
- umap-learn

Machine learning & optimisation
- scikit-learn
- xgboost
- lightgbm
- optuna

Explainability
- shap

Statistics
- scipy
- pingouin

Visualization
- matplotlib
- seaborn

# 12. Annex

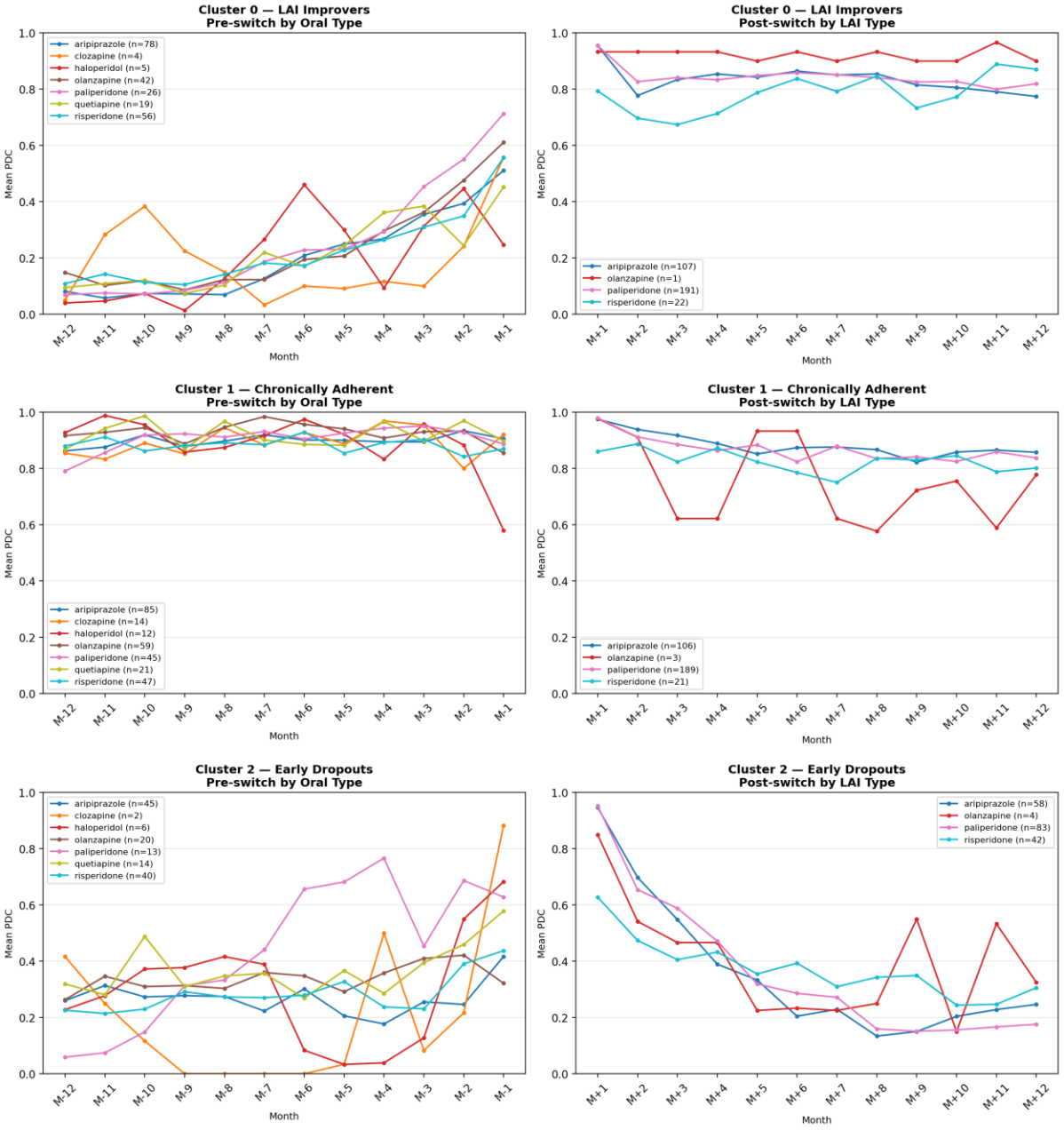
## 12.1. GitHub Repository

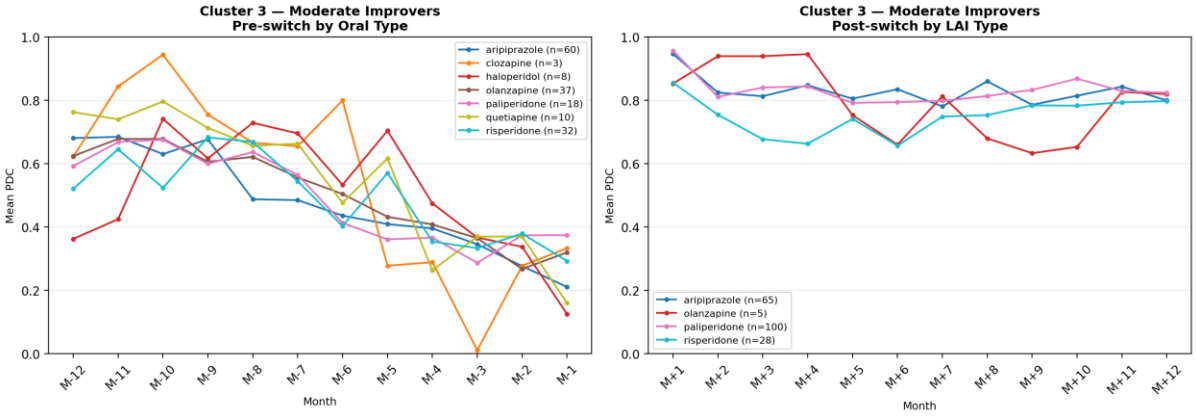
All notebooks were uploaded to a GitHub repository, and PADRIS-PRESTO real data tables were excluded due to legal reasons.

https://github.com/laiaraja/Final_Thesis_Laia.git

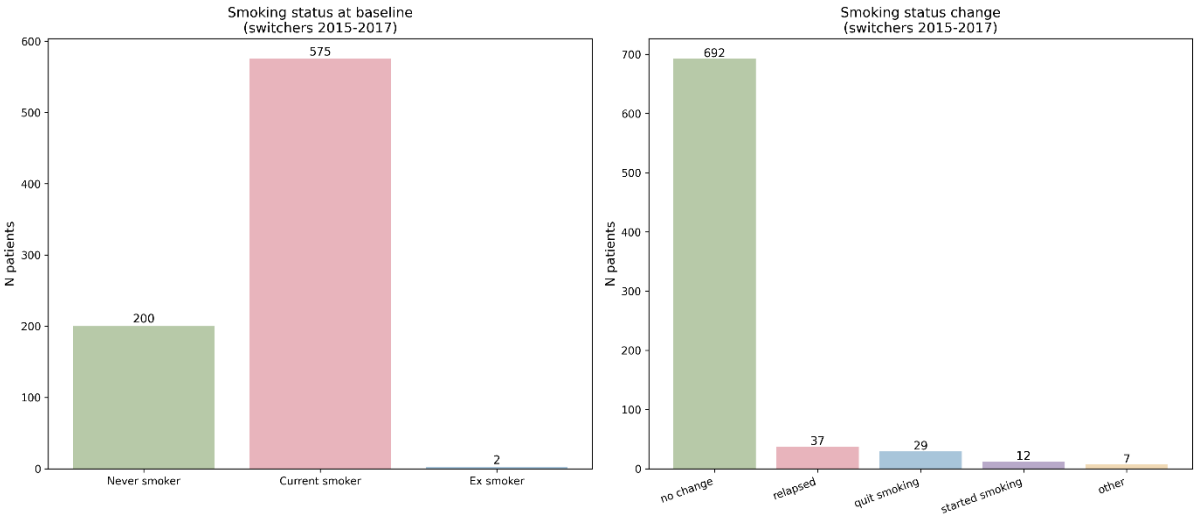
## 12.2. Other graphs

Types of medication per cluster graphs were not considered, although they were commented.

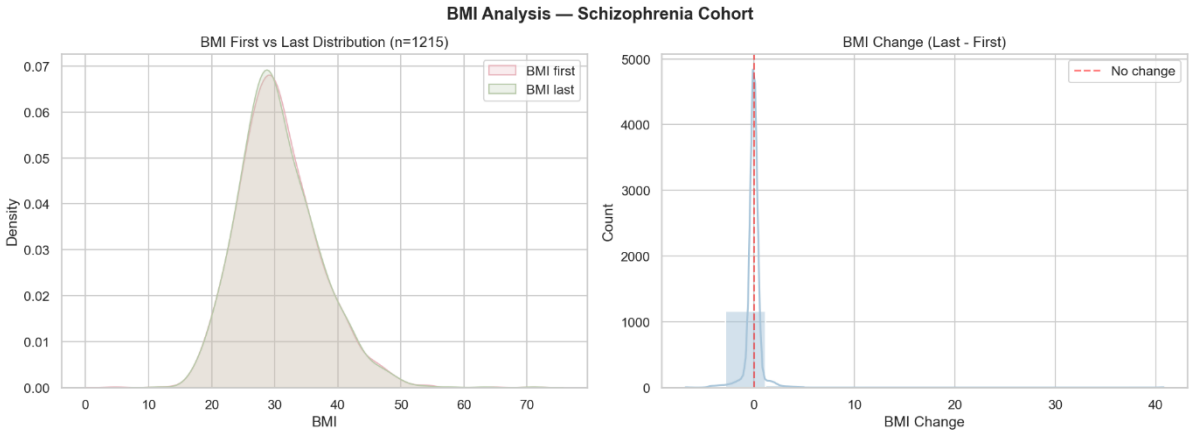


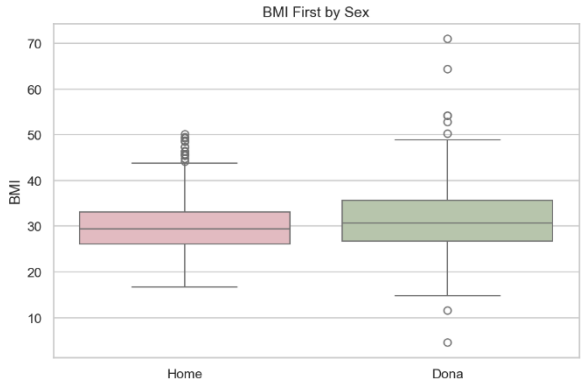


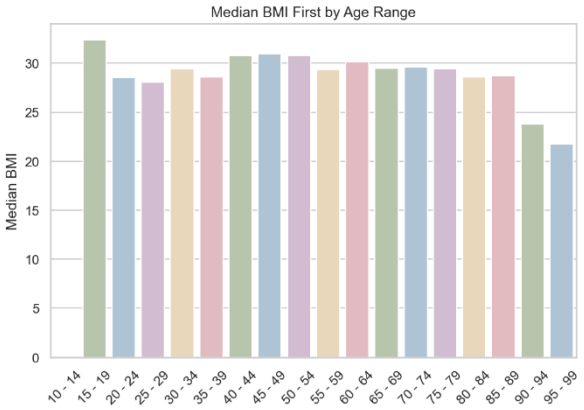

Smoking status at baseline and any changes over the observation period were characterized for the switcher cohort. Smoking data was available for 777 of the 1,025 patients (75.8%). There was no correlation between smoking status changes to schizophrenia diagnosis.




BMI data analysis graphs were not considered as the cohort only included data for the year 2020 which didn't enter the defined cohort.




BMI First by Sex
BMI
70
60
50
40
30
20
10
Home
Dona


Median BMI First by Age Range
Median BMI
30
25
20
15
10
5
0
10 - 14
15 - 19
20 - 24
25 - 29
30 - 34
35 - 39
40 - 44
45 - 49
50 - 54
55 - 59
60 - 64
65 - 69
70 - 74
75 - 79
80 - 84
85 - 89
90 - 94
95 - 99