Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Master Thesis in Computer Engineering

Deep Sequence Modelling of Non-Stationary Disease Progression in Amyotrophic Lateral Sclerosis from Simulated Longitudinal Data

Master Candidate

Filip Trajkovski

Student ID 2110433

Supervisor

Prof. Barbara Di Camillo

University of Padova

Co-supervisor

Prof. Erica Tavazzi

University of Padova

Academic Year
2024/2025

To my family

Abstract

Amyotrophic Lateral Sclerosis (ALS) is a progressive neurodegenerative disease that affects motor neurons in the brain and spinal cord, leading to the gradual loss of voluntary muscle control. Most patients experience rapid functional decline and have a life expectancy of 3 to 5 years from the first symptom's onset. Despite extensive research, there is still no known single cause or curative treatment for ALS. In this context, improving patients' quality of life relies heavily on timely and personalised symptom management. A key step toward this goal is the development of accurate prognostic models capable of predicting individual disease trajectories, enabling clinicians to make informed decisions about interventions and care planning.

Modelling ALS progression presents substantial challenges due to limited real-world longitudinal data, varying visit frequencies, and the inherently non-stationary nature of its evolution as patients transition across disease stages. This thesis addresses these challenges through a hybrid approach that combines probabilistic modelling and deep sequence models. First, we propose a generative modelling approach to simulate realistic patient trajectories using a Dynamic Bayesian Network (DBN) trained on the PRO-ACT dataset which is the largest publicly available ALS clinical trial dataset. The DBN captures temporal dependencies and disease phase transitions, enabling the creation of a large, synthetic cohort with diverse disease trajectories and progression patterns across different clinical phases.

On top of this simulated data, we train and evaluate deep sequence models, including Recurrent Neural Networks such as LSTM and GRU networks, as well as Transformer models with multi-head attention, to predict future values of the ALS Functional Rating Scale subscores. Special attention is given to the impact of non-stationarity on model performance, as disease dynamics shift across phases. We compare the predictive capabilities of these models and their ability to generalise across heterogeneous patient trajectories. The proposed framework offers a scalable and interpretable approach for studying complex, evolving diseases and contributes a reproducible pipeline for synthetic data generation, sequence modelling, and progression prediction.

Sommario

La sclerosi laterale amiotrofica (SLA) è una malattia neurodegenerativa progressiva che colpisce i motoneuroni del cervello e del midollo spinale, provocando una graduale perdita del controllo muscolare volontario. La maggior parte dei pazienti va incontro a un rapido declino funzionale e presenta un' aspettativa di vita di 3-5 anni dall'insorgenza dei sintomi. Nonostante la ricerca, non esistono ancora una causa unica né un trattamento curativo. In questo contesto, il miglioramento della qualità di vita dipende in larga misura da una gestione dei sintomi tempestiva e personalizzata. Un passo fondamentale è lo sviluppo di modelli prognostici accurati, capaci di prevedere le traiettorie individuali della malattia e di supportare decisioni cliniche su interventi e pianificazione delle cure.

Lo studio della progressione della SLA è complesso per la scarsità di dati longitudinali, la variabilità delle visite e la natura non stazionaria dell'evoluzione clinica nelle diverse fasi della malattia. Questa tesi affronta tali sfide con un approccio ibrido che combina modelli probabilistici e di sequenza profonda. In particolare, viene proposta la simulazione di traiettorie realistiche tramite una Rete Bayesiana Dinamica (DBN) addestrata sul dataset PRO-ACT, il più ampio archivio pubblico di studi clinici sulla SLA. La DBN cattura dipendenze temporali e transizioni di fase, permettendo di generare una coorte sintetica ampia ed eterogenea.

Su questi dati vengono addestrati e valutati modelli di sequenza, tra cui reti neurali ricorrenti (LSTM, GRU) e modelli Transformer con meccanismo di attenzione, per predire i futuri sottopunteggi della scala ALSFRS. Particolare attenzione è rivolta all'impatto della non stazionarietà sulle prestazioni, dato il mutare delle dinamiche tra le fasi. Si confrontano le capacità predittive dei modelli e la loro abilità di generalizzare su traiettorie eterogenee. Il framework proposto offre un approccio scalabile e interpretabile allo studio di malattie complesse ed evolutive, contribuendo con una pipeline riproducibile per la generazione di dati sintetici, la modellizzazione di sequenze e la previsione della progressione.

Contents

List of Figures

List of Tables

List of Acronyms

ALS	Amyotrophic Lateral Sclerosis
ALSFRS	ALS Functional Rating Scale
BIC	Bayesian Information Criterion
BN	Bayesian Network
CPT	Conditional Probability Table
DBN	Dynamic Bayesian Network
EM	Expectation-Maximisation
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network

1 Introduction

Modelling disease progression from longitudinal clinical data is one of the central challenges in computational biomedicine. The ability to predict how a disease evolves over time has the potential to improve prognosis of necessary interventions, guide patient management, and optimise the design of clinical trials. However, this task is inherently difficult due to several properties of clinical data. Patient trajectories are highly heterogeneous, reflecting differences in disease onset, demographics, comorbidities, and treatment effects. Missing values are pervasive, arising from skipped visits, recording errors, or inconsistent follow-up schedules. Furthermore, disease processes are rarely stationary as progression rates may change across disease phases, with abrupt accelerations or slowdowns that complicate predictive modelling. These aspects collectively introduce uncertainty that must be explicitly addressed by statistical and machine learning approaches.

Amyotrophic lateral sclerosis (ALS) provides a particularly challenging and clinically relevant case study for disease progression modelling. ALS is a rare, progressive neurodegenerative disease that affects motor neurons, leading to gradual paralysis and ultimately respiratory failure [26]. While the median survival from symptom onset is three to five years, progression is markedly heterogeneous - some patients experience rapid decline, while others progress more slowly [52]. Clinical assessment is typically based on the revised ALS Functional Rating Scale (ALSFRS-R), which measures functional abilities across twelve domains such as speech, swallowing, motor function, and respiration [11]. Although widely adopted in both research and clinical practice, ALSFRS-R trajectories show strong variability between patients, making the prediction of future outcomes highly complex.

From a methodological standpoint, ALS research benefits from the availability of large-scale datasets such as the PRO-ACT database, which aggregates longitudinal data from thousands of patients across multiple clinical trials [4]. This unique resource provides a foundation for both empirical model training and simulation-based research.

Recent advances in machine learning have driven progress in disease progression modelling. Traditional statistical approaches, such as mixed-effects models, remain widely used but are limited in their ability to capture non-linearities and long-range temporal dependencies. Deep learning methods, particularly recurrent neural networks (RNNs) and their variants such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models, have shown strong performance in modelling temporal data across chronic diseases including ALS, Parkinson's, and Alzheimer's disease [50, 24, 57]. These models leverage gating mechanisms to capture complex dependencies while handling irregularly spaced sequences via masking or imputation. More recently, attention-based models and Transformers have been adapted from natural language processing to clinical time series, demonstrating the ability to dynamically identify relevant historical events and better handle variable-length or incomplete data [58]. Survey papers highlight both the promise and the limitations of such models in clinical contexts, emphasising the need for interpretability, robustness, and systematic evaluation under realistic conditions.

Despite these advances, several critical gaps remain in the current literature. Most existing approaches focus on predicting aggregate outcomes, such as total ALSFRS-R scores or survival time, rather than modelling individual functional domains separately [50]. This limits clinical interpretability and reduces the utility of predictions for personalised patient management. Furthermore, many models implicitly assume stationary progression dynamics - that is, they typically learn a single global mapping from input variables to future outcomes, assuming that disease mechanisms remain stable over time. In reality, diseases can exhibit phase-dependent progression, with transitions between early, middle, and late stages characterised by shifts in functional decline rates and symptom correlations. Few studies explicitly address this non-stationarity, and even fewer evaluate model robustness to sudden changes in disease dynamics.

Another limitation is the scarcity of controlled, realistic benchmark datasets for evaluating model performance under different non-stationary conditions. Most published studies rely solely on observational clinical data, which makes it difficult to systematically investigate how models respond to changes in underlying disease dynamics.

This thesis addresses these gaps by focusing on the development of a simulation and model-driven framework for non-stationary disease progression modelling. Using Dynamic Bayesian Networks (DBNs), synthetic longitudinal datasets are generated with controlled perturbations to conditional probability distributions and graph structures, enabling the explicit injection of non-stationarity. These datasets are then used to systematically evaluate deep sequence models, including LSTM, GRU, and Transformer architectures, at multiple levels of granularity, from item-level prediction to phase-specific performance.

The central problem addressed in this thesis can therefore be summarised as follows:

> *How can deep sequence models be designed, trained, and evaluated to accurately predict disease progression while remaining robust to non-stationary dynamics across phases and transitions of the disease?*

By answering this question, we aim to advance the methodological foundations of disease progression modelling and provide insights into how predictive models can be made more adaptive, interpretable, and clinically relevant.

The primary objective of this thesis is to develop and evaluate a framework for modelling non-stationary disease progression using simulated longitudinal data. The focus is on creating datasets that reflect the complexities of real-world progression patterns and on assessing the robustness of deep sequential models under such conditions.

To achieve this overarching goal, the following specific objectives are pursued:

- **Dataset Simulation:** Construct synthetic longitudinal datasets that incorporate realistic disease dynamics. This is achieved through DBNs, which provide a principled probabilistic framework for modelling temporal dependencies while allowing explicit manipulation of non-stationarity.

- **Robustness to Non-Stationarity:** Investigate model performance under non-stationary conditions by introducing controlled perturbations at phase transition points. This allows systematic assessment of each architecture's capacity to adapt to shifts in disease dynamics.

- **Model Development and Training:** Train and evaluate a range of deep sequential architectures, including LSTM, GRU, and Transformer models. These architectures are benchmarked to assess their suitability for progression prediction under complex temporal dynamics.

- **Methodology and Research Design:** Establish a systematic evaluation framework that captures predictive performance at different levels of granularity, including both overall outcomes and functional items, as well as across distinct disease phases and transition points.

The thesis is organised as follows. Chapter 2 provides clinical and methodological background on ALS and progression modelling. Then, Chapter 3 describes the dataset and preprocessing steps. Chapter 4 presents the construction of the DBNs used for simulation, while Chapter 5 outlines the simulation scenarios and generation of synthetic datasets. After that, Chapter 6 introduces the predictive modelling approaches and evaluation framework, while Chapter 7 presents the results across models, scenarios, and datasets. Finally, Chapter 8 concludes with a discussion of findings, implications for ALS research, and future directions.

2

ALS and Models for its Progression

2.1 Overview of ALS

Amyotrophic Lateral Sclerosis (ALS) is a progressive and ultimately fatal neurodegenerative disorder that affects both the upper and lower motor neurons in the brain and spinal cord that control voluntary muscle movement [26]. The degeneration of these neurons results in muscle weakness, spasticity, and eventual paralysis, while cognitive functions are generally preserved in most patients. ALS was first described in 1824 by Charles Bell and later characterised in detail by Jean-Marie Charcot in 1869, who coined the term and established the clinical-pathological correlation still recognised today [46]. Despite major advances in neuroscience, ALS remains incurable, with treatment options limited to symptom management, modest slowing of progression, and maintenance of quality of life.

2.1.1 Epidemiology

ALS affects approximately 5 out of every 100,000 people worldwide, with an annual incidence of 2.16 per 100,000 people in Europe [36]. The disease is slightly more prevalent in men than women and most commonly manifests around the age of 60. The median age at diagnosis is 65.2 years for men and 67 years for women. In the United States, prevalence in 2015 was estimated at 5.2 cases per 100,000 individuals, with higher rates among men (6.4 cases per 100,000 population) and individuals of European ancestry compared to women (4.0 cases per

100,000 population) and non-white populations [37]. Global estimates remain limited by the concentration of studies in Europe and North America, with sparse data available for regions such as Africa, South America, and parts of Asia. The disease course varies considerably between individuals, with median survival estimated at three to five years from symptom onset, although about 10% of patients survive longer than a decade [30]. Familial ALS accounts for approximately 5-10% of cases and is often associated with mutations in genes such as *SOD1* (encoding superoxide dismutase 1), *SETX* (senataxin), *C9orf72*, *TARDBP*, and *FUS/TLS* (Fused in Sarcoma / Translated in LipoSarcoma) [46, 26]. The majority of cases are sporadic, with multifactorial etiologies including environmental exposures and lifestyle factors. Among them, a lifetime of extensive sport or physical exertion, smoking, and active service in the armed forces are related with higher probability of developing the motor neuron disease.

2.1.2 Clinical Course and Pathology

A key clinical challenge in ALS is its heterogeneity in presentation and progression rates. The start of ALS may be so subtle that the symptoms are overlooked. Disease onset is typically classified as spinal (limb-onset) or bulbar, depending on whether initial symptoms affect the limbs or the muscles involved in speech and swallowing. Limb-onset ALS, more common in men, usually begins with weakness in the arms or legs, leading to difficulty with fine motor tasks (such as writing, buttoning a shirt, or turning a key in a lock), tripping or stumbling when walking, or gait disturbances such as foot drop. Bulbar-onset ALS, more common in women and associated with poorer prognosis, manifests initially with dysarthria, dysphagia, or loss of tongue mobility. Less frequently, ALS presents with respiratory-onset, where intercostal muscle weakness leads to breathing difficulties [26].

Over time, symptoms spread regionally, with muscle weakness, atrophy, cramps, fasciculations, significant weight loss, paralysis, dysarthria, and dysphagia becoming more pronounced. Upper motor neuron involvement manifests as spasticity, hyperreflexia, and exaggerated gag reflexes, while lower motor neuron degeneration leads to progressive muscle weakness and atrophy. As the disease advances, most patients lose their ability to walk, speak, or perform activities of daily living independently. Respiratory muscle weakness inevitably develops, with respiratory failure being the most common cause of death, often precipitated and accelerated by pneumonia.

2.1.3 Diagnosis and Prognosis

ALS diagnosis is primarily clinical and based on the revised El Escorial criteria, which require evidence of both upper and lower motor neuron degeneration and symptom spread across multiple body regions [9]. Currently, there is no specific diagnosis test for ALS. This, together with the development of a number of possible non-specific symptoms, explains why the diagnostic delay can be very high, with the median time between symptoms onset and formal diagnosis being 14 months [30]. This fact represents a huge problem, especially because a delayed diagnosis may lead to the use of inappropriate therapies or to the delayed start of fundamental interventions. Differential diagnosis is essential to exclude other neuromuscular conditions such as progressive muscular atrophy, primary lateral sclerosis, multifocal motor neuropathy, myasthenia gravis, or cervical myelopathy [26]. Prognosis is influenced by several factors, including age at onset, progression rate, site of onset, and presence of frontotemporal dementia. Bulbar- and respiratory-onset ALS are generally associated with shorter survival compared to limb-onset disease.

2.1.4 Treatment and Care

There is currently no known cure for ALS. The goal of treatment is to slow the disease progression and improve symptoms. Riluzole and edaravone are the only FDA-approved treatments, with modest benefits in survival and functional progression decline. Symptomatic therapies, such as non-invasive ventilation (NIV), gastrostomy for nutritional support, physiotherapy, and speech therapy, remain central to management and may result in both prolonged survival and improved quality of life [26]. However, they do not stop disease progression. NIV is the primary treatment for respiratory failure in ALS and uses a face or nasal mask connected to a ventilator that provides intermittent positive pressure to support breathing. A feeding tube should be considered if someone with ALS loses 5% or more of their body weight or if they cannot safely swallow food and water. This can take the form of a gastrostomy tube, in which a tube is placed through the wall of the abdomen into the stomach. The feeding tube is usually inserted by a percutaneous endoscopic gastrostomy (PEG) procedure.

It has been demonstrated that ALS patients are better treated in multidisciplinary environment, which can be provided in specialised clinics where psychologists, nutritionists, pneumologists, physical therapists, speech therapists, and specialised nurses cooperate. Despite these advances, the disease trajectory remains relentlessly progressive.

2.1.5 Importance of Predicting Disease Progression

A major challenge in ALS is the marked variability in disease trajectories. Some patients experience rapid decline, while others progress slowly over many years. This heterogeneity complicates prognosis, treatment planning, and the design of clinical trials. Accurate, individualised prediction of disease progression is therefore of substantial clinical importance. Reliable progression models could guide clinicians in tailoring interventions, inform decisions on assistive device provision, and help patients and families prepare for future challenges. For clinical research, such models would improve patient stratification and the evaluation of therapeutic efficacy. The development of robust predictive tools for ALS progression thus represents not only an academic challenge but also a pressing clinical and societal need.

2.2 ALS Functional Rating Scales

In a disease such as ALS, which leads to progressive deterioration of vital capacities including mobility, swallowing, and respiratory function, reliable and user-friendly rating instruments are essential for summarising patients' functional disabilities, correlating with survival prognosis, and monitoring disease progression. Historically, several instruments were used for this purpose, including the Norris Scale [34], the Baylor Appel ALS Rating Scale [3], and the Tufts Quantitative Neuromuscular Examination [38]. Although clinically informative, these scales were often complex and time-consuming, limiting their adoption in both practice and research.

2.2.1 ALS Functional Rating Scale (ALSFRS)

In 1997, the Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS) was introduced as a simpler, patient- and clinician-friendly tool for evaluating

ALS progression [10]. The ALSFRS is a survey administered as a clinical interview or self-reported questionnaire consisting of ten items addressing domains of daily functioning: bulbar tasks (speech, salivation, swallowing), fine motor tasks (handwriting, cutting food, dressing and hygiene), gross motor tasks (turning in bed, walking, climbing stairs), and respiratory function. Each item is scored on a scale from 0 (complete loss of function) to 4 (normal function), yielding a total score between 0 and 40. ALSFRS is easy to administer, sensitive to longitudinal change, and correlates with prognosis and survival. However, it is unbalanced in that only one item directly measures respiratory function, despite respiratory decline being a central determinant of survival.

2.2.2 ALS Functional Rating Scale-Revised (ALSFRS-R)

To address this limitation, Cedarbaum et al.[11] proposed a revised version of the scale in 1999, known as the ALSFRS-Revised (ALSFRS-R). This instrument expands the respiratory domain by adding three specific items: dyspnea, orthopnea, and respiratory insufficiency. The ALSFRS-R thus includes twelve items, each scored from 0 to 4, producing a total score ranging from 0 (maximum disability) to 48 (normal function). The ALSFRS-R has since become the gold standard for functional assessment in ALS due to its reproducibility, sensitivity to change, and prognostic value.

The questions comprising the ALSFRS-R are:

1. Speech
2. Salivation
3. Swallowing
4. Handwriting
5. Cutting food and handling utensils (with/without gastrostomy)
6. Dressing and hygiene
7. Turning in bed and adjusting bed clothes
8. Walking
9. Climbing stairs

10. Dyspnea

11. Orthopnea

12. Respiratory insufficiency

Each item provides clinically meaningful information: for instance, bulbar questions (Q1-Q3) reflect speech and swallowing decline, limb questions (Q4-Q9) capture motor deterioration, and respiratory questions (Q10-Q12) track breathing function. ALSFRS-R score calculated at diagnosis can be compared to scores throughout time to determine the speed of progression. The rate of change, called the ALSFRS-R slope, can be used as a prognostic indicator. Though, the degree of variability is high and a small percentage of people have a much slower progression, on average people with ALS lose about 1 ALSFRS-R point per month. Brief periods of stabilisation ("plateaus") and even small reversals in ALSFRS-R score are not uncommon, due to the fact the tool is subjective and can be affected by medication. However, it is rare for these improvements to be larger (i.e., greater than 4 ALSFRS-R points) or sustained (i.e., greater than 12 months). A survey-based study among clinicians showed that they rated a 20% change in the slope of the ALSFRS-R as being clinically meaningful, which is the most common threshold used to determine whether a new treatment is working in clinical trials.

Despite its strengths, the ALSFRS-R is subject to limitations. It suffers from a "floor effect" in late disease stages, limiting its ability to capture further functional decline, and differences in onset site (bulbar vs. limb) complicate cross-patient comparisons [53]. Consequently, complementary staging systems have been developed.

2.2.3 Staging Systems: King's and MiToS

Two staging systems have been proposed to address limitations of the ALSFRS-R in late disease. The King's staging system, introduced in 2012, defines six clinical stages based on regional spread and major clinical milestones: symptom onset (functional involvement of a first region by weakness, spasticity, dysarthria, or dysphagia), diagnosis, second and third region involvement, need for gastrostomy, and need for respiratory support [45]. The Milano-Torino Staging (MiToS) system defines disease stages by loss of autonomy in four functional domains derived from ALSFRS-R items: walking/self-care, swallowing, communication, and breathing. Each domain is scored as binary (0 = preserved autonomy, 1 = loss of autonomy), with stages determined by the cumulative number of affected domains [16]. While King's staging is particularly useful in early and mid-disease, MiToS is better suited to capturing late-stage progression [22]. However, both systems exclude patients with frontotemporal dementia, limiting their ability to characterise the full ALS phenotype.

2.2.4 Rationale for Item-Level Modelling

Monitoring ALS progression at the item level, rather than relying solely on the ALSFRS-R total score, has gained importance in both clinical and research contexts. Total scores obscure heterogeneous progression patterns across domains: for example, a patient may rapidly lose bulbar function while retaining limb function, or vice versa. Item-level modelling provides a more granular view of disease dynamics, allowing for improved patient stratification, personalised prognosis, and enhanced sensitivity in clinical trials. Moreover, progression of respiratory subscore is closely tied to survival, while bulbar and motor subscores directly impact quality of life and independence [26, 30]. Capturing these trajectories individually is thus essential for understanding ALS heterogeneity and for developing predictive models that account for non-stationary disease progression across phases.

2.3 Challenges in ALS Progression Prediction

Despite advances in clinical characterisation and the development of functional rating scales, the prediction of ALS progression remains a major challenge. Several factors contribute to the difficulty of building reliable models that can generalise across heterogeneous patient populations.

Heterogeneity. ALS exhibits substantial inter-patient and intra-patient variability in its clinical course. Some patients experience rapid functional decline, while others progress more slowly and may survive for over a decade after symptom onset. This heterogeneity spans across multiple dimensions, including age at onset, site of onset, genetic background, and response to treatment. From a modelling perspective, this makes it difficult to define a "typical" disease trajectory and increases the risk of models overfitting to subgroups of patients rather than capturing general progression patterns.

Missingness. Clinical datasets used in ALS research often suffer from missing or incomplete data. Missingness may arise from patients dropping out of studies, skipping follow-up visits, or from variables not being collected at all visits. This introduces bias and complicates the application of standard statistical or machine learning methods. While imputation strategies can partially mitigate this issue, uncertainty remains, and imputed values may propagate errors into the predictive models.

Irregular visit scheduling. In practice, patients are not assessed at uniform time intervals. Some are monitored frequently, while others have long gaps between visits. This irregularity poses challenges for sequence modelling, as traditional models assume equidistant time steps. Techniques such as temporal masking or explicit modelling of time gaps are therefore required to align patient trajectories for analysis.

Non-stationary progression dynamics. A key challenge lies in the fact that ALS progression is not stationary. The rate and pattern of decline may change across different phases of the disease, or shift abruptly around clinical milestones such as the need for respiratory support. Standard models that assume stable dynamics over time may fail to capture such distributional changes. This necessitates the use of flexible modelling approaches that can adapt to evolving temporal patterns and remain robust under phase transitions or structural changes in disease dynamics.

Addressing these limitations is crucial for the development of clinically useful predictive frameworks that can support personalised prognosis and guide therapeutic decision-making.

2.4 Related Work in Disease Progression Modelling

Modelling disease progression from longitudinal clinical data has seen rapid advances in recent years. In particular, there has been growing interest in applying machine learning and deep learning methods to ALS, using large data sources such as PRO-ACT, and explicitly grappling with issues of patient heterogeneity, missing data, and irregular observation schedules.

The PRO-ACT database has been extensively used for ALS progression prediction. For example, [41] applies several neural architectures to predict ALS-FRS slope over months 3-12, comparing deep learning with classical models and showing that deep methods can perform comparably or better in certain settings. Another relevant study [20] leverages PRO-ACT data with demographic, clinical, and laboratory features, comparing XGBoost with Bayesian LSTM (BLSTM) across different observation and prediction windows, while also evaluating model confidence. Similarly, [25] investigates the prediction of clinical outcomes (death, respiratory support, or feeding interventions) over a two-year horizon using routine visit data.

Some recent works explicitly address non-stationary or phase-dependent dynamics. For example, the ML and BLSTM models in [20] implicitly allow for non-stationarity by using past observations across various windows, though they do not explicitly simulate or test robustness to phase shifts. [2] introduces the PASS model, which uses attention mechanisms to allow memoryful, non-Markovian transitions and handles irregular sampling, thereby offering capacity to adapt to changing temporal dynamics. More generally, studies in chronic disease epidemiology have proposed multi-state, non-Markovian frameworks in which past history influences transition probabilities. Such approaches are particularly relevant for modelling non-stationarity in ALS progression.

Some studies have modeled ALS progression at the level of individual functional domains rather than aggregated ALSFRS-R scores. For instance, [51] applies DBNs on multinational registry data to model trajectories for walking, swallowing, breathing, communicating, and survival, enabling sub-cohort simulation and exploration of variable interactions over time. In another line of work, [29] demonstrates that accounting for the multidimensionality of ALSFRS by modelling item-level trajectories or slopes instead of just the total score can improve prediction performance.

More recent efforts have extended the use of PRO-ACT data within deep learning and sequence modelling paradigms. For example, [42] uses PRO-ACT and autoregressive deep architectures to forecast ALS trajectories in various time windos. [56] proposes a Transformer-based architecture trained on longitudinal speech features to predict ALS progression, highlighting how attention-based models can be adapted to clinical time series. Similarly, [44] explores deep learning architectures directly on PRO-ACT, with a focus on handling missingness and optimising prediction performance across different feature sets.

Beyond ALS, research in Alzheimer's disease and other chronic conditions has increasingly adopted attention-based and non-recurrent architectures for progression modelling, including unsupervised trajectory forecasting methods

[2], which are capable of handling irregular sampling and variable temporal dependencies. These studies collectively highlight the growing importance of robustness to non-stationarity and heterogeneity in disease progression modelling.

In summary, while classical statistical and machine learning methods provide strong baselines, recent advances in deep sequence models, particularly recurrent and attention-based architectures, have begun to demonstrate their utility in ALS and related conditions. However, systematic evaluation under controlled non-stationary perturbations remains limited. This gap motivates the simulation-based framework and comparative evaluation of LSTM, GRU, and Transformer models developed in this thesis.

3 Data

3.1 Dataset Overview

The data used in this thesis were obtained from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT), the largest publicly available collection of ALS clinical trial data. The PRO-ACT database was created through a collaboration between ALS Therapy Alliance, Amylyx Pharmaceuticals Inc., Cytokinetics Inc., Knopp Biosciences, Neuraltus Pharmaceuticals Inc., the Neurological Clinical Research Institute at Massachusetts General Hospital, the Northeast ALS Consortium, Novartis, Orion Corporation, Prize4Life Israel, Regeneron Pharmaceuticals Inc., Sanofi, Teva Pharmaceutical Industries Ltd., The ALS Association, and The Sean M. Healey & AMG Centre for ALS at Massachusetts General Hospital [4]. Since its launch in December 2012, more than 3000 researchers from 87 countries (including pharmaceutical companies, informatics companies, and academic institutions) have accessed the database. To date, more than 80 peer-reviewed publications have used PRO-ACT, spanning applications such as: predictive modelling of ALS progression, analysis of epidemiology, identification and verification of prognostic biomarkers, discovery of novel predictors for ALS progression and survival, new methods for stratification of patients based in their baseline characteristics or disease progression profile, simulation of future clinical trials, and development of tools for improved clinical trial recruitment and monitoring [59].

The PRO-ACT initiative brought together patient-level data from 12504 ALS patients who participated in clinical trials sponsored by industry, academic, and foundation partners. It currently contains more than 16 million longitudinally collected data points from patients who participated in 36 phase II and III clinical trials. To protect patient confidentiality, all data are anonymised [4]. Importantly, the database integrates multiple heterogeneous trials into a harmonised format, creating an unprecedented resource for studying disease progression. This also implies that the availability of specific information varies by patient. Within these trials, some patients received placebo treatments, while others received experimental investigational products (IPs). Although none of these IPs demonstrated efficacy compared with placebo, the data still remain invaluable for understanding ALS progression. In a few trials, patients in the treatment arm even appeared to do slightly worse than those in the placebo arm, although whether this reflected drug-specific effects or challenges in patient randomisation remains unclear. For privacy reasons, the names of the specific experimental drugs are not disclosed, but treatment allocation (placebo versus active) is available in the treatment group files.

The dataset is distributed as 16 separate CSV files, each corresponding to a particular type of information (e.g., demographics, functional assessments, laboratory data, or vital signs). Table 3.1 provides an overview of the database files, listing the number of subjects, records, and values contained in each file. Every patient is assigned a unique identifier *SubjectID*, which links records across different data files. Within each file, clinical assessments are timestamped using a variable called *Delta* which represents the number of days since the screening visit (trial baseline is defined as time 0). Negative *Delta* values correspond to events occurring before the trial began, such as symptom onset or prior diagnosis. This consistent temporal alignment makes it possible to reconstruct patient-level trajectories.

The dataset spans a wide variety of clinical, demographic, and biological variables. The **Demographics** file provides essential baseline information on each patient, including age, sex, race, and ethnicity. It also contains age at disease onset, which is a critical prognostic factor, and the time at which demographic data were collected (*Delta* = 0 for screening). These variables serve as anchors for stratifying patients by baseline characteristics and for exploring differences in disease progression across subgroups.

The **Subject ALS History** file captures details of disease onset and diagnosis. It records the site of onset (limb, bulbar, combined, or other), the specific symp-

Table 3.1: Names and characteristics of PRO-ACT database CSV files

Form Name	Subjects	Records	Values
ALL FORMS	12504	3 665 331	17 014 329
Adverse Events	9568	81234	814 583
ALSFRS(R)	8538	73845	985 895
Concomitant Medications	9349	136 072	528 304
Demographics	12504	12504	50815
El Escorial criteria	4243	4243	8485
Family History	1360	1424	5743
Forced Vital Capacity	9090	49110	202 162
Handgrip Strength	1428	19032	172 495
Laboratory Data	10103	2 937 162	11 366 563
Mortality	5038	5043	8579
Muscle Strength	1522	204 875	1 840 381
Riluzole use	10253	10363	20696
Slow Vital Capacity	4294	21115	103 435
Subject ALS History	11100	13765	44646
Treatment Group	10823	10823	18367
Vital Signs	11075	84721	843 180

toms reported, and the anatomical locations where symptoms first appeared. It also contains the time intervals between symptom onset, diagnosis, and trial entry (*Onset Delta* and *Diagnosis Delta*). This information provides valuable insights into diagnostic delays, which can be predictive of disease severity and allows modelling of disease duration prior to trial enrollment.

Closely related is the **El Escorial Criteria** file, which documents diagnostic classifications based on established clinical guidelines. Categories include "Definite", "Probable", "Possible", and "Suspected" ALS, each recorded with its *Delta* timestamp. These criteria provide an additional layer of diagnostic precision and enable evaluation of how disease certainty evolves over time.

The **Family History** file records whether close relatives were affected by ALS or other neurological disorders. It provides a detailed breakdown by relative type (e.g., father, mother, sibling, grandparent, etc.), as well as genetic information when available. Although available for only a subset of patients, this file is particularly important for distinguishing sporadic from familial ALS cases, which may follow different clinical trajectories.

The **ALSFRS(R)** file is among the most clinically significant. It contains assessments from both the original ALSFRS and its revised version (ALSFRS-R).

The original ALSFRS has 10 items scored from 0 (no function) to 4 (normal function), while the ALSFRS-R expands this to 12 items by subdividing respiratory function into three separate questions. Item-level scores, total scores, and *Delta* values for each assessment are included. These measures are the gold standard for tracking ALS progression and are widely used in both research and clinical settings. In those cases where a question score was missing, but scores were available for that question from previous and subsequent measures, the score was imputed by the original data donors, by averaging the two measures. This sometimes results in non-integer scores.

Complementing functional scores are the **Forced Vital Capacity (FVC)** and **Slow Vital Capacity (SVC)** files, which measure respiratory function. FVC records the volume of air forcibly blown out after full inspiration, while SVC captures the volume of air that can be exhaled slowly after slow maximum inhalation. SVC is typically greater than FVC. Both are expressed in litres and as a percentage of predicted normal values adjusted for age, sex, and height. Since respiratory decline is a hallmark of ALS and a strong predictor of mortality, these measures provide essential endpoints for survival and disease progression models.

The **Laboratory Data** file is one of the most extensive, containing thousands of variables representing blood and urine test results. Examples include complete blood counts, electrolyte levels, liver enzymes, kidney function markers, glucose, cholesterol, and immune parameters. Each entry is associated with a test name, result, unit, and *Delta*. Although subject to variability across trials, this file enables exploratory biomarker analyses and may reveal subtle physiological patterns related to ALS progression.

The **Vital Signs** file includes repeated measures of blood pressure (systolic and diastolic, supine and standing), pulse, respiratory rate, body temperature, height, and weight. Given the importance of weight loss and malnutrition in ALS prognosis, these longitudinal measurements are particularly valuable for survival modelling and progression prediction.

The **Muscle Strength** and **Hand Grip Strength** files provide quantitative assessments of motor function. Muscle strength is measured across multiple muscle groups (e.g., ankle dorsiflexion, elbow flexion, hip flexion) using hand-held dynamometry, while grip strength specifically evaluates hand function. These objective measures complement the more subjective ALSFRS scores and allow for more granular analysis of motor decline.

The **Treatment Group** file specifies whether patients received placebo or an IP and records the timing of treatment initiation relative to trial baseline. While the specific drugs tested are not disclosed, the dataset notes whether each patient was in an active or placebo arm. The **Riluzole Use** file separately tracks use of riluzole, the only widely approved ALS treatment during the period of data collection. These files together provide insights into treatment exposure and allow comparisons between arms, even though none of the IPs showed overall efficacy.

The **Concomitant Medications** file records other drugs taken by patients during the trial, including dose, frequency, and timing. These medications may have been used to manage symptoms, comorbidities, or adverse events. Because concomitant medications were not always consistently reported, absence of data does not necessarily indicate lack of use.

The **Adverse Events** file documents all medical events recorded during the trials, ranging from mild symptoms such as headaches to severe conditions such as stroke. Events are coded hierarchically, from specific terms to broader categories, and include information on severity, outcome, and timing.

Finally, the **Mortality** file provides survival outcomes. It records whether each patient died during follow-up and, if so, the number of days from trial entry to death. This file is crucial for survival modelling, risk stratification, and validation of prognostic models.

The breadth of information contained in PRO-ACT makes it a uniquely valuable dataset for ALS research. By pooling data from multiple trials, it offers statistical power that far exceeds any individual study, enabling more robust investigations of disease progression, prognostic modelling, and biomarker discovery. Its longitudinal nature is particularly well-suited for machine learning methods that require temporal data. Furthermore, the availability of both clinical outcomes (ALSFRS decline, survival) and physiological measures (e.g., lung function, laboratory biomarkers) supports integrative approaches that can link functional deterioration to underlying biological changes.

Despite its strengths, PRO-ACT also presents challenges typical of real-world clinical trial data. Measurement protocols vary across studies, missing values are frequent, and patients contribute different numbers of follow-up visits. These issues require careful preprocessing and methodological consideration. Nevertheless, the PRO-ACT database has already catalysed numerous advances in ALS research and continues to serve as a cornerstone resource for developing and evaluating predictive models of disease progression [references].

3.2 Ethical and Privacy Considerations

The use of clinical trial data for secondary research purposes requires careful attention to ethical and privacy considerations. The PRO-ACT database was explicitly designed to enable broad sharing of patient-level information while ensuring the highest standards of patient confidentiality. To achieve this, the curators of PRO-ACT implemented strict anonymisation and compliance procedures in accordance with the Health Insurance Portability and Accountability Act (HIPAA) de-identification conventions for personal health information (PHI). All patient identifiers, including names, addresses, contact details, and any other directly identifying information, were removed prior to data release. Dates were shifted relative to each patients trial baseline (*Delta* = 0), and variables that could indirectly lead to reidentification were carefully reviewed and standardised to minimise disclosure risk.

Because the trials were originally conducted under informed consent and approved by institutional review boards, patient rights and ethical considerations were already safeguarded at the point of data collection. However, the additional step of harmonising and anonymising the datasets was critical in enabling public access without compromising individual privacy. The PRO-ACT initiative followed internationally recognised data sharing principles, ensuring that the released dataset could be used by researchers, clinicians, and data scientists worldwide while remaining compliant with ethical and legal requirements.

Importantly, PRO-ACT data are only accessible through a controlled access process. Researchers are required to register on the platform, agree to the terms and conditions of use, and confirm that they will not attempt to reidentify participants. This governance framework reflects best practices in data management, balancing the benefits of open scientific collaboration with the responsibility to protect vulnerable patient populations.

By adhering to these procedures, the PRO-ACT database has established itself as a benchmark resource for ethical large-scale data sharing in neurodegenerative disease research. Its success demonstrates that it is possible to provide open access to rich, individual-level clinical trial data while fully preserving patient privacy and complying with relevant ethical and legal frameworks.

3.3 Data Preprocessing

The preprocessing pipeline used in this thesis builds directly on prior work that transformed the raw PRO-ACT dataset into a unified, analysis-ready format [23]. Although the structure of the PRO-ACT files was relatively consistent, the fragmentation across multiple sources posed significant challenges for downstream analysis, particularly for machine learning models that require integrated longitudinal data. The primary goal of the earlier preprocessing phase was therefore to merge these heterogeneous files into a single, harmonised dataset while ensuring quality, consistency, and interpretability.

The workflow began with a systematic audit of missingness and redundancy. First, the columns that contained only zeros or missing values were identified and removed. This step eliminated variables with no informative content, reducing noise in the dataset. The removed columns varied by file. For example, some administrative metadata in the ALSFRS(R) file and certain family history descriptors were excluded, while other files such as Adverse Events, Demographics, and Laboratory Data were retained in full.

Once the irrelevant columns had been pruned, the available information was categorised into three major groups based on temporal behaviour, as indicated by the variable *Delta*:

1. **Static variables**, which remain constant across all visits (e.g., demographics, family history, treatment arm).
2. **Dynamic variables**, which vary over time (e.g., ALSFRS scores, vital signs, laboratory values).
3. **Temporary variables**, which occur within a defined temporal window (e.g., adverse events, concomitant medications).

This classification provided a conceptual framework for subsequent processing, ensuring that static variables were not inadvertently treated as longitudinal and that temporary variables were handled separately.

Each file was then carefully reviewed and transformed. In the **Demographics** file, redundant time columns such as *Demographics_Delta* were dropped, binary race variables were consolidated into a single categorical column, and inconsistent notations (e.g., "Americ_Indian_Alaska_Native" vs. "Native American") were harmonised. Similarly, ethnicity categories were standardised, and implausible codes were recoded as missing. No age outliers were detected, confirming that values lay within the plausible clinical range of 18 to 88 years.

The **Subject ALS History** file required more extensive restructuring. Because some individuals reported multiple symptoms or onset sites, the data were reformatted to allow for multiple entries per patient (e.g., `Symptom_1`, `Symptom_2`, `Symptom_3` with associated locations). Duplicates without additional information were removed, inconsistent encodings were unified (e.g., "both" recoded to "limb and bulbar"), and static variables such as *Subject_ALS_History__Delta* were eliminated.

In the **Family History** file, numerous binary indicators for different relatives were collapsed into a single categorical column indicating which family members were affected by ALS or other neurological conditions. This step improved interpretability and reduced sparsity.

The **Death** file underwent standardisation of column names and values, ensuring consistent coding of censored and uncensored survival times. Where survival status was ambiguous (e.g., missing but with recorded visits), censored status was imputed using the date of the last available observation.

For the **El Escorial Criteria**, which classifies diagnostic certainty, only baseline assessments were retained (as the variable is static), and redundant time variables were removed.

Both **Forced Vital Capacity** and **Slow Vital Capacity** files were standardised by renaming columns, removing redundant unit fields, and averaging duplicate measurements recorded on the same day. Rows with missing *Delta* were excluded.

The **Vital Signs** file required several unit conversions: heights from inches to centimetres, weights from pounds to kilograms, and temperatures erroneously recorded in Fahrenheit converted to Celsius. Duplicates were averaged, implausible entries were corrected or removed, and static measurements such as height were fixed across all visits for each patient.

The **ALSFRS(R)** file, which contains the primary functional outcome measures, was also carefully cleaned. Missing or inconsistent *Delta* values were removed, duplicates averaged, and overlapping items (e.g., Q5a vs. Q5b for with and without gastrostomy feeding) were merged. Total scores were recalculated when inconsistent with items, and ALSFRS Q10 items were reconstructed from ALSFRS-R respiratory items where necessary. This step was particularly important because ALSFRS items later served as the modelling targets.

The **Laboratory Data** file was the most complex and underwent extensive cleaning. Non-informative rows (e.g., missing test results, missing *Delta*) were removed, test names and units were harmonised, and implausible values were flagged and corrected. Some tests that were recorded too infrequently (fewer than ten times across the entire dataset) were excluded, while others were split into multiple variables to account for different units of measurement (e.g., urine protein measured in dipstick scale vs. mg/dL vs. mg/24h). Outliers with physiologically impossible values were discarded. Duplicates were resolved by averaging, and red blood cell morphology information was consolidated into a single string variable.

After these domain-specific preprocessing steps, all static and dynamic variables were merged into a single unified dataset, leaving aside temporary variables such as adverse events and concomitant medications, which were not the focus of progression modelling [23].

A final round of cleaning was then performed. Variables incorrectly typed as character but representing numeric values were converted, monotonic properties of ALSFRS scores were checked, and derived features such as MiToS functional domains (movement, swallowing, communication, breathing) were computed from ALSFRS items. Visits recorded before baseline (negative *Delta*) were excluded, and subjects with implausible onset dates or insufficient follow-up (fewer than two visits in the first three months and one visit at six months) were removed. Importantly, patients from both placebo and active treatment arms were retained, since none of the IPs demonstrated efficacy in the original trials.

Finally, missingness was quantified at the patient level rather than the record level, reflecting whether a given variable was ever measured for a patient within the first three months. Variables missing in more than 50% of patients were excluded. After this filtering, the dataset consisted of **5441 patients, 61 variables, and 69,166 total observations**.

The following part details the additional preprocessing steps carried out in this thesis, building on the already preprocessed dataset provided in [23] and described before. The overarching goals were to obtain complete and reliable target labels (ALSFRS items), harmonise patient-level covariates, handle missingness with principled imputation under plausibility constraints, and produce a compact, discretised feature set suitable for DBNs.

3.3.1 Feature Removal

Several variables were excluded due to redundancy, irrelevance, or sparsity. Race was removed because the distribution was overwhelmingly dominated by a single category (Caucasian), which limited the predictive value for functional progression and left little variability for modelling. Variables such as *Treatment Group Delta* and *Riluzole use Delta* had near-zero values for most patients, offering minimal temporal signal. The MiToS composite domains (Movement, Swallowing, Communicating, Breathing) were excluded as they were redundant, summarising ALSFRS items already modelled individually. Finally, *Subj_condition* and *Time_Event* were removed, as they were primarily relevant to mortality modelling rather than functional progression. In addition, the *Onset Site* variable was harmonised by merging "Spine" and "Limb" categories into a unified limb-onset group, as they represent clinically equivalent onset forms.

3.3.2 Imputation of ALSFRS items

The prediction targets were defined as the ten ALSFRS items, each measured on a scale from 0 to 4. Because reliable targets are essential for supervised learning, missingness was addressed deterministically where possible, following clinically logical rules.

1. **Total Score Recovery:** If all ten items were observed but the ALSFRS Total score was missing, the total was recalculated as their sum.

2. **Item Recovery:** If exactly one item was missing but the ALSFRS Total score was available, the missing item was inferred as the difference. This was only accepted if the inferred value lay within the valid range [0, 4].

Cases where this inference produced implausible values were left unchanged. This procedure successfully recovered 1828 individual item values across multiple variables. However, in 254 cases the inferred values fell outside the valid range and were therefore discarded. The majority of these invalid inferences occurred for `Q5_Cutting`. Table 3.2 summarises the initial number of missing values, the number of recovered values, the count of invalid inferences, and the final number of missing values for each ALSFRS variable.

Rows with unresolved ALSFRS items after this conservative step were subsequently dropped, ensuring complete, reliable and valid targets for supervised

Table 3.2: Summary of deterministic inference of missing ALSFRS items. A total of 1828 values were recovered, while 254 invalid inferences (outside the valid 0-4 range) were discarded.

Item	Initial Missing	Recovered	Invalid Inferences	Final Missing
ALSFRS Total	18742	0	0	18742
Q1 (Speech)	13448	0	0	13448
Q2 (Salivation)	13450	3	0	13447
Q3 (Swallowing)	13451	0	1	13451
Q4 (Handwriting)	13452	3	0	13449
Q5 (Cutting)	20796	1809	249	18987
Q6 (Dressing and Hygiene)	13453	0	1	13453
Q7 (Turning in Bed)	13456	4	1	13452
Q8 (Walking)	13454	2	0	13452
Q9 (Climbing Stairs)	13452	0	1	13452
Q10 (Respiratory)	13457	7	0	13450
Total	156611	1828	254	156037

learning. Overall, the deterministic, range-checked recovery of ALSFRS values avoids training on ambiguous labels while salvaging rows where the logic is sound.

3.3.3 Patient Filtering

To guarantee meaningful longitudinal sequences, patients with fewer than five visits were excluded. This decision reflected two considerations:

1. Simulated data later in the thesis assumes at least five visits per patient.

2. Sequence models require adequate temporal context to capture progression dynamics. Very short series provide limited and insufficient signal, leading to unstable training and risk of biasing models toward short-horizon behaviour.

After applying the filtering criteria, the dataset was reduced from 5441 to 4819 patients as shown in Table 3.3. This filtering step ensured that only patients with sufficiently long longitudinal follow-up were included in the analyses. The remaining dataset consisted of 49143 visit records across 52 variables.

The number of visits per patient after filtering showed a mean of approximately 10.2 visits, with a standard deviation of 3.2, as it can be seen in Table 3.4.

Table 3.3: Patient and visit counts before and after filtering based on missing ALSFRS values and minimum number of visits set to 5.

	Patients	Visits
Before filtering	5441	69166
After filtering	4819	49143

Table 3.4: Descriptive statistics for the number of visits per patient before and after filtering.

Statistic	Mean	Std	Min	25%	Median	75%	Max
Before filtering	12.71	6.23	3	9	12	13	100
After filtering	10.20	3.19	5	7	11	12	22

3.3.4 Handling of Missing Data and Imputation

Missing values are a pervasive issue in clinical datasets, and the PRO-ACT database is no exception. Laboratory results, vital signs, and demographic descriptors are frequently incomplete due to skipped tests, recording errors, or inconsistent follow-up schedules. To ensure robust modelling, it is essential to address missingness in a principled manner.

The distribution of categorical variables after filtering is reported in Table 3.5, where it can be seen that several categorical descriptors had missing values. To address this, plausible defaults were assigned to balance completeness with conservative assumptions. Specifically, missing Riluzole use entries were set to "No" based on the assumption that treatment initiation would typically be recorded explicitly. Study Arm missing values were defaulted to "Active" under the assumption that placebo assignment would generally be specified. Sex missing values were imputed as "Male" reflecting dataset demographics where the majority of the patients are male. For Onset Site, missing values were categorised as "Other" to avoid unfounded assumptions. The resulting distributions after imputation are shown in Table 3.6.

In the end, categorical variables were transformed into numerical representations using label encoding. The encoding scheme for each categorical variable is summarised in Table 3.7.

Continuous variables with more than 50% missingness were excluded, as their imputation would introduce excessive uncertainty. Specifically, three hematology counts were removed: Absolute Lymphocyte Count, Absolute

Table 3.5: Descriptive statistics of categorical variables after filtering. Frequencies and proportions are reported across all observations (N = 49143).

Variable	**Category**	**Frequency (%)**
Sex	Male	31259 (63.6%)
	Female	17884 (36.4%)
Onset Site	Limb	34178 (69.5%)
	Bulbar	9716 (19.8%)
	Other	4728 (9.6%)
	Limb and Bulbar	509 (1.0%)
	Missing	12 (0.0%)
Riluzole Usage	Yes	28031 (57.1%)
	No	9972 (20.3%)
	Missing	11140 (22.7%)
Study Arm	Active	27872 (56.7%)
	Placebo	19557 (39.8%)
	Missing	1714 (3.5%)

Table 3.6: Descriptive statistics of categorical variables after filtering and imputation of missing data. Frequencies and proportions are reported across all observations (N = 49143).

Variable	**Category**	**Frequency (%)**
Sex	Male	31259 (63.6%)
	Female	17884 (36.4%)
Onset Site	Limb	34178 (69.5%)
	Bulbar	9716 (19.8%)
	Other	4740 (9.6%)
	Limb and Bulbar	509 (1.0%)
Riluzole Usage	Yes	28031 (57.1%)
	No	21112 (42.9%)
Study Arm	Active	29586 (60.2%)
	Placebo	19557 (39.8%)

Table 3.7: Label encoding scheme applied to categorical variables.

Variable	**Encoding Scheme**
Sex	0 = Female, 1 = Male
Onset site	0 = Bulbar, 1 = Limb, 2 = Limb and Bulbar, 3 = Other
Subject used Riluzole	0 = No, 1 = Yes
Study Arm	0 = Active, 1 = Placebo

Monocyte Count, and Absolute Neutrophil Count (Table 3.8).

Table 3.8: Variables with more than 50% missingness that were excluded from the dataset.

Variable	Missingness (%)
Absolute Monocyte Count (10^9/L)	51.78
Absolute Lymphocyte Count (10^9/L)	51.77
Absolute Neutrophil Count (10^9/L)	51.77

The descriptive statistics of the continuous numerical features after filtering are summarised in Table 3.9, including the number of available and missing values, as well as the main distributional characteristics (minimum, quartiles, median, and maximum).

In this thesis, missing data are imputed using the *Iterative Imputer* [43], a method conceptually related to the Multiple Imputation by Chained Equations (MICE) algorithm.

The statistical treatment of missing data is often guided by the mechanism of missingness described by [47]. Three main mechanisms are recognised:

- **Missing Completely at Random (MCAR):** the probability of missingness is independent of both observed and unobserved data.

- **Missing at Random (MAR):** the probability of missingness depends only on observed data.

- **Missing Not at Random (MNAR):** the probability of missingness depends on unobserved values themselves.

The Iterative Imputer, like most modern imputation methods, assumes that data are at least MAR. Under this assumption, missing values can be reasonably predicted from the information available in observed covariates.

The MICE framework addresses missingness by specifying a conditional distribution for each variable with missing entries [54]. Chained regression models are then fitted to approximate these conditional distributions, and missing values are updated iteratively. Conceptually, MICE treats the data as a multivariate distribution:

$$P(X_1, X_2, \ldots, X_p)$$

.

Table 3.9: Descriptive statistics of continuous numerical variables after filtering. Each patient contributes with all their observations (total N = 49143).

Variable	Count	Missing	Min	25%	50%	75%	Max
Age	47170	1973	18	47	56	64	84
Onset Delta	48724	419	-8741	-852	-563	-371	-16
Delta	49143	0	0	62	154	288	2114
FVC (L)	34388	14755	0	2.18	3.01	3.94	7.68
Absolute Eosinophil Count (10^9/L)	29203	19940	0	0.10	0.16	0.32	30
Albumin (g/L)	33449	15694	18	40	43	46	59
Alkaline Phosphatase (U/L)	32990	16153	1	62	75	90	1705
ALT (U/L)	36041	130102	1	21	29	41	944
AST (U/L)	35956	13187	0	21	27	34	911
Bicarbonate (mmol/L)	31711	17432	2.5	24	26	28	46
Bilirubin (μmol/L)	34996	14147	0	6.84	10	15	153.94
Blood Urea Nitrogen (mmol/L)	36245	12898	0.73	4.38	5.40	6.50	18.70
Calcium (mmol/L)	33447	15696	0.25	2.25	2.35	2.44	5.91
Chloride (mmol/L)	34670	14473	75	101	103	105	135
CK (U/L)	31486	17657	5	117	211	373	22232
Creatinine (μmol/L)	36256	12887	0	53.04	69	88.4	360
Glucose (mmol/L)	34815	14328	1.17	4.8	5.28	5.94	31.5
Hematocrit (%)	34273	14870	0.29	40	43	45.3	81
Hemoglobin (g/L)	34453	14690	11.4	136	145	153	280
Platelets (10^9/L)	33378	15765	0.000192	200	238	281	866
Potassium (mmol/L)	36026	12917	2	4	4.1	4.4	42
Protein (g/L)	32989	16154	42	69	72	75	96
Red Blood Cells (10^9/L)	31616	17527	0.044	4.3	4.7	5	40
Sodium (mmol/L)	36079	13064	88	139	140	142	182
White Blood Cells (10^9/L)	31636	17507	1	5.8	6.9	8	21.9
Absolute Basophil Count (10^9/L)	26503	22640	0	0.02	0.03	0.06	7590
Gamma-glutamyltransferase (U/L)	26407	22736	3	18	26	40	1646
Blood Pressure - Diastolic (mmHg)	29466	19677	20	74	80	88	140
Blood Pressure - Systolic (mmHg)	29473	19670	70	120	130	140	220
Height (cm)	36348	12795	131	163	170.2	178	205
Pulse (beats/minute)	29568	19575	18	69	76	84	144
Respiratory Rate (breaths/minute)	34610	14533	6	16	18	20	90
Weight (kg)	36750	12393	28.5	62.6	72.7	83.4	189

When some entries are missing, the framework approximates this distribution by learning conditional models:

$$P(X_j \mid X_{-j}),$$

where X_{-j} denotes all variables except X_j. Each missing entry is then imputed from these models, and the process is repeated for all variables in a cycle until convergence.

Formally, the Iterative Imputer operates as follows:

1. **Initialisation:** Replace missing entries with initial guesses, typically column means or medians.

2. **Iterative cycle:** For each feature with missing values, partition the data into the target column X_j and predictors X_{-j}. Fit a regression model using the observed entries of X_j and predict the missing entries from X_{-j}. Update the dataset with these predictions.

3. **Convergence:** Repeat the cycle until the imputed values stabilise or until a fixed maximum number of iterations is reached.

The Bayesian Ridge Regression model for imputation was selected in this thesis because it is particularly well-suited for continuous clinical variables: it handles multicollinearity effectively and applies shrinkage to coefficients, reducing overfitting. Logistic regression or decision tree models can be used for categorical variables, but in this case categorical features were label-encoded as integers, allowing them to be handled numerically by the imputer.

Several settings were used to ensure stability and plausibility of imputations:

- A maximum of 100 iterations was allowed to ensure convergence.

- The initial strategy was mean imputation, providing a simple starting point.

- Feature-wise minimum and maximum values were enforced, computed directly from the observed data. This ensured that imputed values remained physiologically plausible and consistent with the clinical ranges present in the dataset.

Iterative imputation has several advantages that made it particularly suitable for this work. Because each variable is modelled conditionally on all others, correlations between features are preserved. This is especially important in ALS data, where relationships between measures such as FVC, respiratory rate, and ALSFRS items reflect meaningful clinical dependencies. Moreover, enforcing empirical bounds on imputations prevented implausible values, such as negative laboratory measures or physiologically impossible vital signs.

In summary, iterative imputation was chosen because simpler methods such as mean or median substitution would fail to preserve correlations and could distort relationships critical for both DBNs and deep learning models. By modelling each variable as a function of the others and iteratively refining estimates until convergence, the Iterative Imputer reconstructed missing values in a way that preserved multivariate structure, respected biological plausibility, and integrated seamlessly into the preprocessing pipeline.

3.3.5 Enforcing Patient-Level Consistency

Certain features are inherently static (i.e., time invariant) and should remain constant across visits. These include sex, age at onset, onset site, onset delta, study arm, and Riluzole exposure. However, intra-patient inconsistencies were observed, likely arising from data entry errors, harmonisation mismatches or imputation errors across trials. To enforce consistency, variables such as age at onset and onset delta were replaced with the rounded patient-level mean across visits. This correction ensured intra-patient stability in static features, a critical requirement for DBNs and deep learning models, as small inconsistencies could otherwise corrupt temporal models with confusing data.

3.3.6 Derived Features

A clinically meaningful temporal axis called *Time Since Onset* was created as:

$$\text{Time Since Onset} = |\text{Onset Delta}| + \text{Delta} \tag{3.1}$$

This measure aligns all visits on a clinically intuitive clock starting from the real disease onset rather than from the diagnosis point or study baseline as they can differ among patients. Based on this, three discrete disease phases were defined: 0-2 years, 2-4 years, and 4+ years since onset, encoded as discretised

Time Since Onset $\in \{1, 2, 3\}$. Table 3.10 summarises the distribution of patients across the three phases of disease progression. The majority of patients remain within a single phase or transition between two phases, while only a very small number experience trajectories spanning all three phases.

Table 3.10: Distribution of patients across Time Since Onset phases.

Phase trajectory	Number of patients
Only in phase 1	1087
Only in phase 2	871
Only in phase 3	293
In both phases 1 and 2	2100
In both phases 2 and 3	462
In all phases 1, 2, and 3	6

Table 3.11 reports the distribution of visits per patient across the three phases. While patients in all phases show similar central tendencies, phase 1 patients tend to have slightly more visits on average, whereas variability in visit counts is higher in phase 2 and phase 3.

Table 3.11: Visit statistics per *Time Since Onset* phase.

Phase	Patients	Min visits	Max visits	Mean visits	Median visits	Std
1	3193	1	18	7.11	7	3.30
2	3439	1	21	6.27	6	3.83
3	761	1	13	6.41	6	3.94

Body Mass Index (BMI) was derived from height and weight as follows:

$$\text{BMI} = \frac{\text{Weight (kg)}}{\left(\frac{\text{Height (cm)}}{100}\right)^2} \tag{3.2}$$

In this study, BMI was treated as a dynamic feature, since it varies over time as the patient's weight changes. At each clinical visit, BMI was calculated based on the recorded or imputed weight (in cases where weight measurements were missing), while height was assumed constant for each patient. This ensured that BMI values were consistently available across all visits, capturing temporal changes in nutritional status, which are relevant for ALS progression. After computing BMI, the raw variables Height (cm) and Weight (kg) were dropped to reduce redundancy.

Table 3.12 summarises the descriptive statistics of the continuous numerical features after the previously performed preprocessing steps. These statistics provide an overview of variable distributions and highlight the heterogeneity present in the ALS population.

Table 3.12: Descriptive statistics of continuous numerical features after imputation.

Feature	Mean	Std	Min	25%	50%	75%	Max
Delta	188.49	159.69	0.00	62.00	154.00	288.00	2114.00
FVC (L)	3.08	1.15	0.00	2.25	3.07	3.88	7.68
Absolute Eosinophil Count (10^9/L)	0.48	1.17	0.00	0.00	0.11	0.31	30.00
Albumin (g/L)	43.24	3.05	18.00	41.00	43.03	45.00	59.00
Alkaline Phosphatase (U/L)	77.46	24.68	1.00	65.00	75.00	85.00	1705.00
ALT (U/L)	33.45	18.86	1.00	23.00	31.00	38.43	944.00
AST (U/L)	29.37	12.46	0.00	23.00	28.00	32.00	911.00
Bicarbonate (mmol/L)	26.11	2.92	2.50	24.40	26.00	27.64	46.00
Bilirubin (μmol/L)	10.31	5.50	0.73	6.84	9.37	12.83	153.94
Blood Urea Nitrogen (mmol/L)	5.60	1.45	0.25	4.75	5.50	6.24	18.70
Calcium (mmol/L)	2.36	0.11	0.25	2.28	2.35	2.43	5.91
Chloride (mmol/L)	102.76	3.04	75.00	101.43	103.00	104.00	135.00
CK (U/L)	297.31	258.36	5.00	145.00	251.00	379.00	22232.00
Creatinine (μmol/L)	67.74	17.99	0.00	55.58	68.32	79.56	360.00
Glucose (mmol/L)	5.55	1.37	1.17	4.88	5.32	5.80	31.50
Hematocrit (%)	34.46	17.02	0.29	37.80	41.90	44.70	81.00
Hemoglobin (g/L)	145.00	11.65	11.40	137.00	146.01	152.00	280.00
Platelets (10^9/L)	243.90	64.15	0.00	213.00	245.68	275.87	866.00
Potassium (mmol/L)	4.19	0.39	2.40	4.00	4.18	4.30	42.00
Protein (g/L)	72.06	3.92	44.00	70.00	71.98	74.00	96.00
Red Blood Cells (10^9/L)	11.84	14.37	0.00	4.37	4.80	5.35	40.00
Sodium (mmol/L)	140.32	2.39	88.00	139.00	140.29	141.00	182.00
White Blood Cells (10^9/L)	10.09	6.15	1.00	6.30	7.16	10.00	21.90
Absolute Basophil Count (10^9/L)	0.64	0.48	0.00	0.00	0.00	1.00	3.00
Gamma-glutamyltransferase (U/L)	55.04	11.63	0.00	47.00	56.00	64.00	7590.00
Blood Pressure - Diastolic (mmHg)	46.29	50.62	3.00	22.00	32.00	50.43	1646.00
Blood Pressure - Systolic (mmHg)	78.97	10.89	20.00	74.00	80.00	84.00	220.00
Pulse (beats/min)	130.13	13.76	70.00	121.12	130.00	136.98	205.00
Respiratory Rate (breaths/min)	79.29	12.09	18.00	72.00	78.00	84.58	144.00
Age	55.04	11.63	18.00	47.00	56.00	64.00	84.00
BMI	25.32	4.01	10.97	22.91	25.31	26.90	55.82

3.3.7 Discretisation of Variables

For DBN compatibility and interpretability, continuous variables were discretised. A hybrid strategy was employed. For selected laboratory markers with simple, clinically meaningful thresholds as Absolute Eosinophil Count and Absolute Basophil Count (thresholded at $0.5 \times 10^9/L$), binary indicators were created. The other continuous numerical features were discretised into tertiles using the thresholds shown in Table 3.13, corresponding to the 33rd and 67th percentiles, essentially dividing their values into low (0), medium (1), and high (2) categories. This is done in order to achieve compact conditional probability tables (CPTs) for DBNs in the next step.

3.3.8 Final Dataset Assembly

After all preprocessing and transformations, the final dataset consisted of:

- **Static variables**: Subject ID, Sex, Age, Onset Site, Study Arm, Riluzole treatment.
- **Temporal variable**: Time Since Onset (discretised).
- **Dynamic variables**: discretised laboratory and vital features, including:
  - **Hematological and biochemical markers:** Eosinophil, Basophil, Albumin, Alkaline Phosphatase, ALT, AST, Bicarbonate, Bilirubin, Blood Urea Nitrogen, Calcium, Chloride, CK, Creatinine, Glucose, Hematocrit, Hemoglobin, Platelets, Potassium, Protein, Red Blood Cells, Sodium, White Blood Cells, Gamma-glutamyltransferase.
  - **Respiratory and cardiovascular measures:** FVC, Diastolic Blood Pressure, Systolic Blood Pressure, Pulse, Respiratory Rate, BMI.
- **Targets (ALSFRS items)**:
  - Q1: Speech
  - Q2: Salivation
  - Q3: Swallowing
  - Q4: Handwriting
  - Q5: Cutting food and handling utensils
  - Q6: Dressing and hygiene
  - Q7: Turning in bed and adjusting bed clothes
  - Q8: Walking
  - Q9: Climbing stairs

Table 3.13: Tertile cut-off values for discretised continuous numerical features. Values represent the thresholds at the 33rd and 67th percentiles used for discretisation. For Absolute Eosinophil Count and Absolute Basophil Count, clinical thresholds were applied instead of tertiles.

Feature	**33rd Percentile**	**67th Percentile**
Age	50.0	61.0
FVC (L)	2.54	3.60
Albumin (g/L)	42.0	44.41
Alkaline Phosphatase (U/L)	68.91	81.0
ALT (U/L)	26.0	36.03
AST (U/L)	25.0	31.0
Bicarbonate (mmol/L)	25.0	27.0
Bilirubin (μmol/L)	7.90	11.89
Blood Urea Nitrogen (mmol/L)	5.03	6.04
Calcium (mmol/L)	2.31	2.40
Chloride (mmol/L)	102.0	103.97
CK (U/L)	177.99	337.69
Creatinine (μmol/L)	61.88	74.81
Glucose (mmol/L)	5.05	5.63
Hematocrit (%)	39.71	43.90
Hemoglobin (g/L)	140.0	150.15
Platelets (10^9/L)	225.0	264.0
Potassium (mmol/L)	4.10	4.23
Protein (g/L)	70.40	73.0
Red Blood Cells (10^9/L)	4.5	5.0
Sodium (mmol/L)	140.0	141.0
White Blood Cells (10^9/L)	6.70	8.50
Gamma-glutamyltransferase (U/L)	25.0	41.0
Blood Pressure - Diastolic (mmHg)	77.0	82.89
Blood Pressure - Systolic (mmHg)	125.24	134.05
Pulse (beats/min)	74.68	81.96
Respiratory Rate (breaths/min)	16.67	19.89
BMI	23.91	26.26
Clinically Thresholded Features	**Criterion**	**Binary Encoding**
Absolute Eosinophil Count (10^9/L)	0.5	1 = Above, 0 = Below
Absolute Basophil Count (10^9/L)	0.5	1 = Above, 0 = Below

- Q10: Respiratory function

Data were sorted by *Subject ID* and visit order to preserve original sequence order, with *Delta* dropped since temporal order and *Time Since Onset were explicitly modelled*. For DBN structure and parameter learning, a flattened dataset was also created, pairing consecutive visits into time steps $t-1$ and t.

4 Dynamic Bayesian Network

4.1 Theoretical Overview

Dynamic Bayesian Networks (DBNs) are a class of probabilistic graphical models specifically designed to represent stochastic processes that evolve over time [33]. They extend the principles of static Bayesian Networks (BNs), which capture conditional dependencies among variables at a single time slice, to temporal domains by providing an explicit representation of sequential dynamics. In this way, DBNs offer a compact and interpretable framework for modelling complex longitudinal systems in which current states depend not only on contemporaneous variables but also on their historical trajectories [13].

A BN is a directed acyclic graph (DAG) whose nodes represent random variables and whose edges represent direct probabilistic dependencies between variables. Each node is associated with a conditional probability distribution that quantifies the effect of its parents in the graph [32]. Mathematically, a BN defines a joint probability distribution over all variables $X = \{X_1, X_2, \ldots, X_n\}$ as

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i)),$$

where $Pa(X_i)$ denotes the set of parent nodes of X_i in the DAG. This factorisation allows efficient representation and inference in high-dimensional spaces when conditional interdependence is present.

BNs have been widely applied in domains such as diagnosis, decision support, and prediction [31, 60]. However, they are inherently static, representing dependencies at a single snapshot in time, and thus cannot directly capture the dynamics of processes that evolve. A DBN generalises a BN to model sequences of variables over time [39]. Instead of representing a single set of variables, a DBN considers a sequence of time steps $t = 1, 2, \ldots, T$, each containing a set of variables $X^{(t)} = \{X_1^{(t)}, \ldots, X_n^{(t)}\}$.

Formally, a DBN consists of two primary components:

- **Intra-slice dependencies**, which define conditional relationships among variables within the same time step.

- **Inter-slice dependencies**, which capture temporal dependencies by linking variables at time t to those at time $t + 1$.

Through this factorisation, DBNs efficiently model joint probability distributions over sequences of random variables. A common assumption, also adopted in this thesis, is the *first-order Markov property* [18], which states that the future state depends only on the present state and not directly on the past beyond it:

$$P\left(X^{(t+1)} \mid X^{(1)}, \ldots, X^{(t)}\right) = P\left(X^{(t+1)} \mid X^{(t)}\right).$$

This leads to the two-slice temporal BN representation, in which the DBN is specified by:

- **Initial network** $P(X^{(1)})$: a BN describing dependencies at the initial time step.

- **Transition network** $P(X^{(t+1)} \mid X^{(t)})$: a BN describing how variables evolve from one time slice to the next.

The joint distribution over the entire sequence is then given by:

$$P\left(X^{(1)}, \ldots, X^{(T)}\right) = P(X^{(1)}) \prod_{t=2}^{T} P\left(X^{(t)} \mid X^{(t-1)}\right).$$

This simplification allows for tractable inference and simulation of longitudinal processes, as the same transition structure and parameters can be "unrolled" over an arbitrary number of time steps.

4.1.1 Structure Learning

The process of learning a DBN begins with the identification of the network's structure, i.e., the set of edges that encode conditional dependencies among variables. Structure learning can be performed using either *constraint-based* or *score-based* approaches [31]. The constraint-based approach assumes that the network topology is specified *a priori* based on domain knowledge in the form of known causal relationships, while in the score-based approach a search over possible graph structures is performed with the goal of optimising a scoring function.

In this thesis, a score-based approach was employed, in which candidate network structures are iteratively evaluated according to a goodness-of-fit criterion. This enables the model to adapt to the statistical relationships present in the dataset. Specifically, the Bayesian Information Criterion (BIC) was used, as it balances model fit against structural complexity, penalising overly dense networks that risk overfitting [40]. The BIC score of a given network structure is defined as:

$$\text{BIC} = \log L(\hat{\theta} \mid D, G) - \frac{k}{2} \log n,$$

where:

- $L(\hat{\theta} \mid D, G)$ is the likelihood of the data D given the maximum likelihood estimates $\hat{\theta}$ of the parameters for the network structure G,
- k is the number of free parameters in the model,
- n is the number of data points.

The first term rewards models that explain the observed data well, while the second term penalises complexity by increasing with the number of parameters. In DBN structure learning, the BIC score is computed for each candidate network, and the structure with the highest BIC score is preferred. This approach is advantageous in medical data modelling, where overly complex networks may capture noise rather than true causal relationships.

The Hill-Climbing algorithm was utilised to perform the search over candidate structures. Hill-Climbing is an iterative, greedy optimisation procedure that starts from an initial structure and explores local modifications such as adding, removing, or reversing edges [14]. At each step, the modification yielding the greatest improvement in the chosen score (BIC) is accepted, and the process continues until no further improvements are possible. Although Hill-Climbing is a heuristic method and does not guarantee global optimality, it is widely applied in practice due to its efficiency and its ability to yield clinically interpretable structures. This interpretability arises because the algorithm typically converges to sparse networks that reflect plausible conditional dependencies between variables rather than overly complex or dense graphs. In the context of clinical data, such networks are valuable as they can highlight relationships consistent with medical knowledge, while avoiding spurious connections that would obscure meaningful patterns.

4.1.2 Parameter Learning

Once the structure of the DBN has been defined, parameter learning is performed to estimate the Conditional Probability Tables (CPTs) associated with each variable. Each CPT specifies the probability distribution of a variable given the states of its parent variables. When the data are fully observed, parameter learning reduces to computing maximum-likelihood estimates of conditional probabilities. In practice, however, clinical datasets frequently contain missing or partially observed values, making direct estimation infeasible. In such cases, the Expectation-Maximisation (EM) algorithm is commonly employed, as it allows parameter estimation in the presence of missing data by iteratively alternating between estimating missing values based on current parameters (E-step) and updating parameters given the completed data (M-step), converging to locally optimal parameter estimates [19, 32]. In this work, however, EM was not required since all missing values were addressed through preprocessing and imputation prior to DBN learning, ensuring that parameter estimation could be performed directly on fully observed data.

4.1.3 Application to Patient Simulation

The adoption of DBNs in this thesis is motivated by their suitability for modelling disease progression in ALS [60]. ALS is a heterogeneous and dynamic condition, where patient trajectories differ substantially in both rate and pattern of decline. DBNs provide a principled means of representing these dynamics by:

- capturing temporal dependencies in ALSFRS items, ensuring that future values are informed by prior states,

- allowing incorporation of clinical constraints, such as monotonicity in functional decline, through the structure and conditional distributions,
- enabling simulation of synthetic patient trajectories by sampling from the learned model, with the flexibility to inject controlled forms of non-stationarity (e.g. phase-specific changes in CPTs).

Through these properties, DBNs serve as a powerful generative framework for producing realistic ALS progression datasets. Such simulated data not only mitigate the scarcity of longitudinal clinical trial data but also provide a controlled environment for evaluating the robustness of deep sequence models under diverse non-stationary scenarios.

4.2 Construction of the DBN

To generate realistic synthetic ALS patient trajectories, a DBN was constructed using the preprocessed dataset introduced in Chapter 3. The objective is to create a probabilistic graphical model that accurately reflects clinically meaningful dependencies across demographic, baseline, and dynamic variables, while also maintaining computational tractability for large-scale simulation.

To enforce temporal ordering and biologically plausible causal relationships, variables were organised into six conceptual layers:

1. **Demographics:** sex.
   (a) Can influence: Layers 2, 3, and 6.
   (b) Cannot be influenced by other layers.
2. **Onset Characteristics:** age at onset and onset site.
   (a) Can influence: Layers 2 (self-loops, i.e., dependencies between onset variables), 3, and 6.
   (b) Can be influenced by: Layers 1 and 2.
3. **Treatment Factors:** study arm and Riluzole treatment.
   (a) Can influence: Layer 6 (dynamic variables at time t).

   (b) Can be influenced by: Layers 1 and 2.

4. **Time Since Onset:** time since onset at visit $t-1$.
   (a) Can influence: Layer 6.
   (b) Cannot be influenced by other layers (serves as an exogenous input).

5. **Dynamic Variables at Visit $t-1$:** laboratory values, vital signs, and ALSFRS items.
   (a) Can influence: Layer 6.
   (b) Cannot be influenced by other layers.

6. **Dynamic Variables at Visit t:** the same set of variables as in Layer 5, measured at the subsequent visit.
   (a) Cannot influence earlier layers (ensures forward temporal flow).
   (b) Can be influenced by: Layers 1, 2, 3, 4, and 5.

The layering strategy, including categories of variables, their membership to layers, and permissible directional relationships, is summarised in Table 4.1.

Table 4.1: Layering strategy of the DBN, showing categories of variables, their assignment to layers, and permitted directional influences between them.

Layer	Category	Variables	Can Influence	Can Be Influenced By
1	Demographics	Sex	Layers 2, 3, 6	None
2	Onset Characteristics	Age at onset, Onset site	Layers 2, 3, 6	Layers 1, 2
3	Treatment Factors	Study arm, Riluzole	Layer 6	Layers 1, 2
4	Time Since Onset	Time since onset at $t-1$	Layer 6	None
5	Dynamic Variables ($t-1$)	Vitals, labs, ALSFRS items	Layer 6	None
6	Dynamic Variables (t)	Vitals, labs, ALSFRS items	None	Layers 1-5

This design ensures that static demographic and onset characteristics shape downstream disease evolution and may influence treatment allocation, while dynamic variables evolve under the dual influence of prior states ($t-1$) and the passage of time.

Mandatory edges were enforced from *Time Since Onset* (Layer 4) to all dynamic variables at visit t (Layer 6) in order to ensure that disease duration drives clinical outcomes. This reflects the fundamental clinical principle that ALS progression is strongly dependent on elapsed time from disease onset.

Finally, to maintain computational feasibility and avoid overfitting, a maximum of five parents per node was imposed. This cap ensures that CPTs remain interpretable and estimable despite the discretised multi-level variable space.

5 Data Simulations

5.1 Simulation Scenarios

The baseline DBN described in Section 4.2 provides a generative foundation for modelling ALS progression trajectories under the assumption of *stationarity*, i.e., that the conditional distributions governing transitions between functional states remain constant over time. However, real-world disease progression rarely exhibits such stationary dynamics. In particular, ALS trajectories are often marked by heterogeneity, phase-dependent progression, sudden accelerations of decline, and increased variability [15].

To simulate these phenomena, we designed a series of non-stationary perturbation scenarios by systematically altering the CPTs or the network structure across phases. Each scenario represents a distinct hypothesis about the mechanisms of ALS progression, translated into mathematical transformations of the underlying distributions. The goals of these scenarios are twofold:

1. To mimic clinically relevant phenomena, such as rapid decline, uncertainty in late phases, or restricted reversibility.

2. To stress-test sequential models (LSTM, GRU, Transformer) under systematically perturbed distributions.

Non-stationarity was introduced by altering the CPTs associated with target variables, specifically the ALSFRS items at time t. Each transformation modified the conditional distributions of functional decline while respecting clinical constraints, such as the unlikelihood of long-term functional improvement.

Formally, if a CPT specifies a conditional distribution

$$p(y_t \mid x_{t-1}), \tag{5.1}$$

then a transformation $T(\cdot)$ is applied such that

$$\tilde{p}(y_t \mid x_{t-1}) = T\big(p(y_t \mid x_{t-1})\big), \tag{5.2}$$

where $\tilde{p}$ denotes the perturbed probability distribution. The key property of the transformation is that the probabilities remain valid, i.e.

$$\sum_y \tilde{p}(y \mid x_{t-1}) = 1. \tag{5.3}$$

Several classes of perturbations were implemented, each designed to capture a different mode of non-stationarity. These included flattening of distributions to simulate increased uncertainty, skewing toward decline to simulate rapid progressors, temperature scaling to adjust entropy, step-size banding to enforce local continuity, variance inflation to model heteroscedasticity, and structural modifications to reflect evolving causal dependencies. Each of these scenarios is described in detail in the following subsections.

5.1.1 Scenario 1: CPT Flattening

The first non-stationary scenario considered is based on the idea of *flattening* the CPTs of the DBN in a way that reflects greater uncertainty and variability in patient trajectories as the disease progresses. In the original formulation of the DBN, the conditional probabilities $p(y_t \mid x_{t-1}) = \{p_1, p_2, \ldots, p_K\}$ describe the likelihood of observing a given functional score at time t given the state at the previous time step. These distributions can flexibly encode different transitions across the phases of the disease.

Clinical intuition and empirical observations both suggest that in the early stages of the disease, progression is usually gradual and predictable. However, such determinism weakens over time, and transitions become more biased toward decline. As patients advance into later phases of ALS, their disease trajectories become increasingly diverse, influenced by secondary complications, treatment adherence, and individual variability.

To capture this clinical reality, the flattening procedure introduces a decline-based reference distribution $b = \{b_1, b_2, \ldots, b_K\}$, which places greater weight on lower ALSFRS values, corresponding to functional decline. Formally, this distribution is defined as

$$b_i = \frac{K - i + 1}{\sum_{j=1}^{K} j} = \frac{2(K - i + 1)}{K(K + 1)}, \quad i = 1, \ldots, K, \tag{5.4}$$

where K denotes the number of possible states of the functional score and i indexes the individual state, ordered from worst ($i = 1$) to best ($i = K$) functional outcome.

The transformation of the original distribution into a flattened distribution is then achieved through a convex combination of the two distributions:

$$\tilde{p} = \alpha p + (1 - \alpha) b, \quad \alpha \in [0, 1]. \tag{5.5}$$

Here, the parameter α acts as a phase-specific mixing coefficient. Larger values of α correspond to trajectories that remain closer to the original dynamics, while smaller values place more weight on the decline-biased distribution. Figure 5.1 compares the original probability distribution with the flattened one when parameter $\alpha = 0.8$.

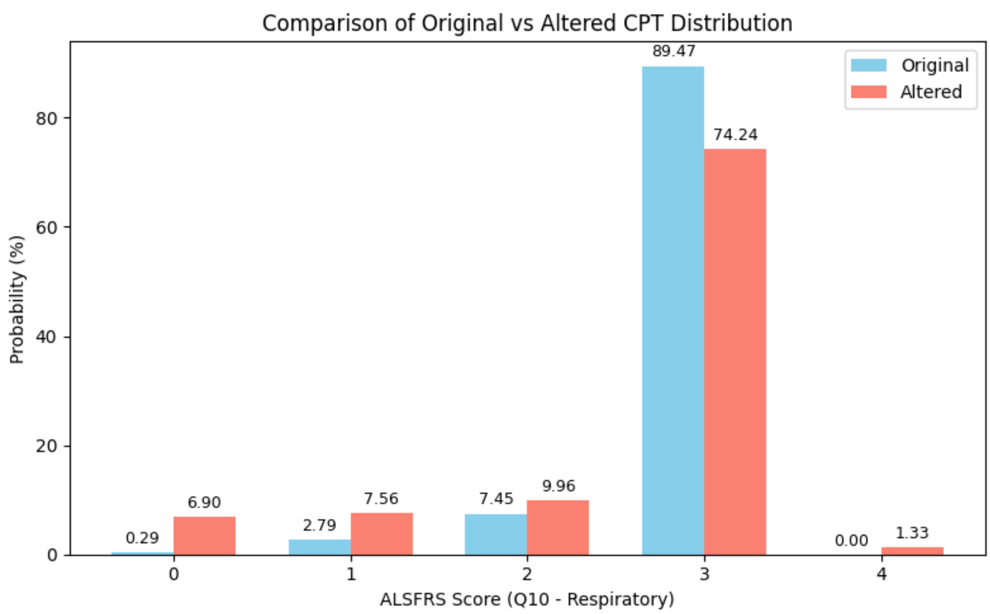

Figure 5.1: Comparison of the original and altered CPT distributions for `Q10 - Respiratory`, using CPT Flattening with $\alpha = 0.8$.

5.1.2 Scenario 2: Extreme Skewing

The second non-stationary scenario introduces *extreme skewing* of the conditional probability distributions, designed to mimic the clinical observation that in later phases of ALS, patients are substantially more likely to deteriorate than remain stable or improve. Unlike the flattening transformation in Scenario 1, which primarily increased uncertainty while maintaining some balance across potential outcomes, skewing explicitly redistributes probability mass toward worse functional states.

Mathematically, the skewing procedure relies on a weight vector that increases the probability of lower values of the ALSFRS items. For each functional domain with K possible states, a deterioration-oriented weighting scheme is defined as

$$w_i = (K - i + 1)^{\gamma}, \quad \gamma \geq 1, \quad i = 1, \ldots, K, \tag{5.6}$$

where w_i assigns larger weights to smaller values of i, corresponding to more severe levels of impairment. The exponent γ acts as a skew strength parameter, determining how strongly the distribution is biased toward decline. When $\gamma = 0$, the weighting reduces to unity, leaving the original probabilities unchanged. When $\gamma = 1$, the bias grows linearly, and as γ becomes large, the transformation increasingly concentrates probability mass on the lowest functional states, approaching an almost deterministic collapse into severe impairment.

The original conditional probability distribution $p = \{p_1, p_2, \ldots, p_K\}$ is then modified according to the weighted normalisation:

$$\tilde{p}_i = \frac{p_i w_i}{\sum_{j=1}^{K} p_j w_j}, \quad i = 1, \ldots, K, \tag{5.7}$$

resulting in a new distribution $\tilde{p}$ that explicitly favors deterioration while preserving valid probabilistic structure. Figure 5.2 illustrates the comparison of the original probability distribution with the altered one when parameter $\gamma = 2$.

5.1.3 Scenario 3: Temperature Scaling

A third non-stationary scenario leverages *temperature scaling* as a means of systematically adjusting the confidence or entropy of conditional probability distributions without altering their underlying support. This transformation is motivated by the clinical observation that the progression of ALS is not uni-

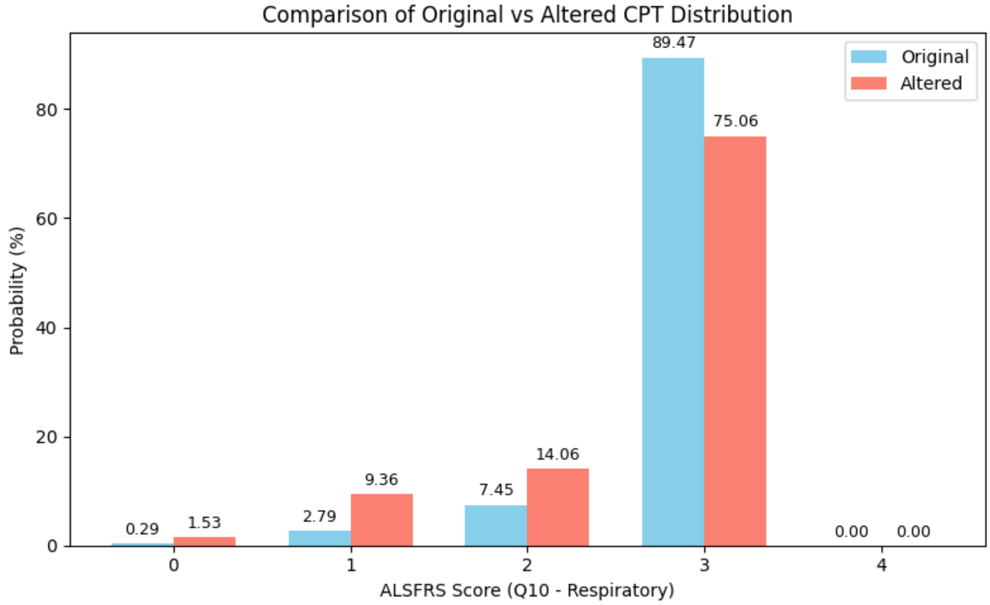

Figure 5.2: Comparison of the original and altered CPT distributions for `Q10 - Respiratory`, using Extreme Skewing with $\gamma = 2$.

formly deterministic: in some phases, outcomes appear highly uncertain and heterogeneous across patients, while in others, decline follows a relatively predictable pattern. Temperature scaling offers a mathematically principled mechanism to capture such variability in outcome uncertainty, allowing us to explore how predictive models respond when the entropy of disease dynamics changes across phases.

Formally, the transformation introduces a temperature parameter $T > 0$, which acts as a scaling factor on the conditional probability distribution $p = \{p_1, p_2, \ldots, p_K\}$. The modified distribution is given by

$$\tilde{p}_i = \frac{p_i^{1/T}}{\sum_{j=1}^{K} p_j^{1/T}}, \quad i = 1, \ldots, K. \tag{5.8}$$

The role of the temperature parameter is to soften or sharpen the probability distribution. For values $T > 1$, the transformation produces a flatter distribution, corresponding to greater uncertainty or noise in the transition dynamics, thereby increasing the entropy of the system. Conversely, for $T < 1$, the transformation sharpens the distribution, concentrating probability mass on the most likely outcome and reducing entropy, effectively simulating more deterministic disease progression. Figure 5.3 compares the original probability distribution with the altered one when $T = 1.2$.

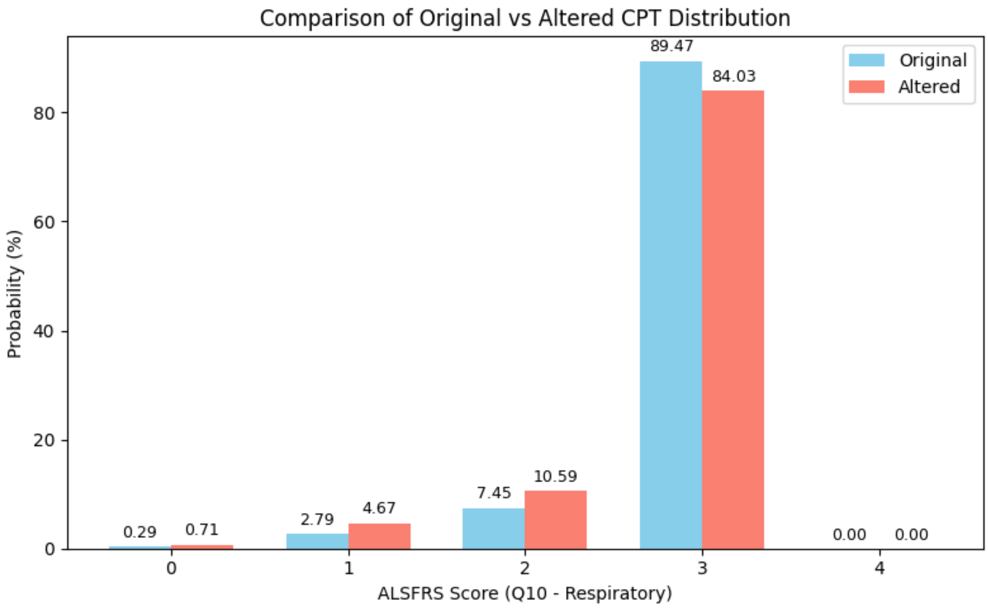

Figure 5.3: Comparison of the original and altered CPT distributions for `Q10 - Respiratory`, using Temperature Scaling with $T = 1.2$.

5.1.4 Scenario 4: Step-size Banding

A fourth scenario introduced *step-size banding*, a transformation designed to enforce local continuity in disease progression trajectories by constraining the size of permissible transitions in ALSFRS items between consecutive visits. This scenario is motivated by a fundamental clinical reality: ALS is a gradually progressive neurodegenerative disorder, and although patients may differ in their rate of decline, it is exceedingly rare to observe large, abrupt jumps in functional ability between two visits spaced a few months apart. For instance, a patients speech or walking ability typically does not collapse from near-normal function to complete loss within a single assessment interval. Instead, deterioration tends to occur in small, incremental steps. Similarly, substantial functional recovery is biologically implausible given the irreversible nature of motor neuron degeneration. The purpose of this transformation is thus to ensure that simulated patient trajectories respect local continuity, while still allowing phase-specific variation in the likelihood of decline.

Mathematically, the step-size banding approach modifies the conditional probability distribution of each ALSFRS item by introducing a masking function that zeroes out transitions exceeding a maximum allowed step size. Let $s_{t-1} \in \{1, \dots, K\}$ denote the state of an item at the previous time step and let $p = \{p_1, p_2, \dots, p_K\}$ denote the corresponding conditional distribution for the next state s_t. A step-size parameter $J \geq 0$ specifies the largest allowable change between consecutive states. The step-size mask is defined as

$$m_i = \begin{cases} 1 & \text{if } |i - s_{t-1}| \leq J, \\ 0 & \text{otherwise,} \end{cases} \qquad i = 1, \dots, K. \tag{5.9}$$

The masked distribution is then obtained by elementwise multiplication followed by renormalisation:

$$q_i = \frac{p_i m_i}{\sum_{j=1}^{K} p_j m_j}, \quad i = 1, \dots, K. \tag{5.10}$$

To avoid strictly zero probabilities and maintain numerical stability, a small smoothing parameter $\varepsilon \in [0, 1]$ is optionally introduced, yielding the smoothed distribution:

$$\tilde{p}_i = (1 - \varepsilon) q_i + \varepsilon \frac{1}{K}, \quad i = 1, \dots, K. \tag{5.11}$$

This formulation ensures that only transitions within the allowable band $\pm J$ of the current state retain most of the probability mass, while large and clinically implausible jumps are effectively eliminated. The smoothing step guarantees that no state is entirely excluded, thereby preventing hard constraints from producing brittle dynamics. Figure 5.4 compares how the original probability distribution is changed when the parameters are set to $J = 1$ and $\epsilon = 0.05$.

Moreover, step-size banding can be combined with other transformations, such as temperature scaling, to create more realistic and challenging scenarios. For instance, one might first soften the distribution using a high temperature to introduce uncertainty, and then apply banding to enforce biological plausibility, thereby simulating patients whose progression is noisy but still locally smooth.

5.1.5 Scenario 5: Variance Inflation

The fifth non-stationary scenario introduces *variance inflation*, a probabilistic transformation designed to inject controlled uncertainty into the CPTs of the DBN. The underlying idea is to reduce the sharpness of the distributions by

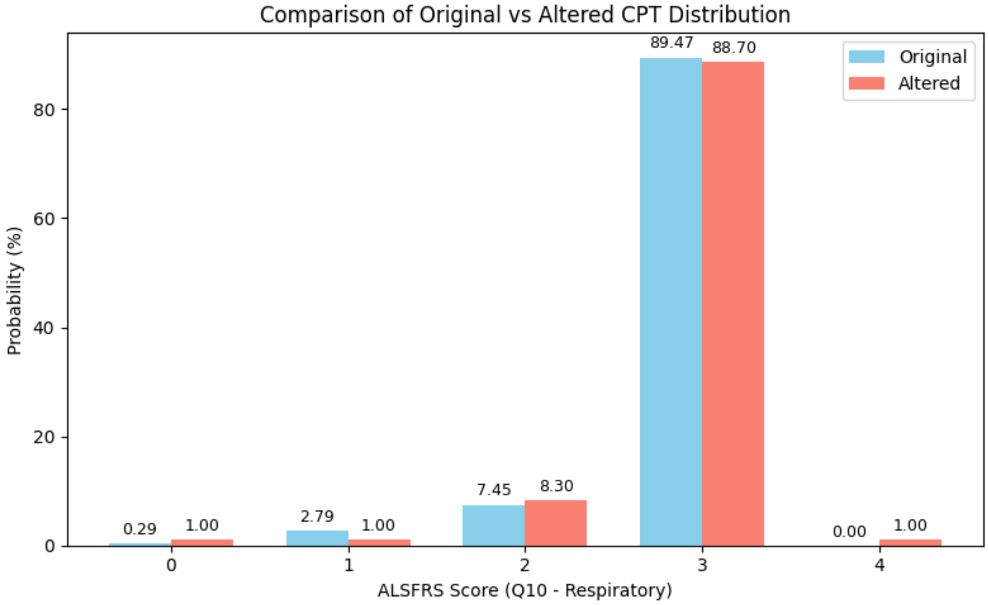

Figure 5.4: Comparison of the original and altered CPT distributions for `Q10 - Respiratory`, using Step-size Banding with $J = 1$ and $\epsilon = 0.05$.

redistributing part of the probability mass from the most likely outcome (the mode) to its immediate neighbours. This reflects the clinical observation that, while ALS progression generally follows a declining path, the precise rate of decline is not perfectly predictable as patients may deviate slightly from their expected course due to biological variability, comorbidities, or treatment-related effects. By increasing the variance of the CPTs while preserving their local structure, this transformation prevents artificially deterministic trajectories and yields synthetic data that better mirrors the uncertainty present in real-world ALS progression.

Formally, consider an original conditional distribution for a given ALSFRS item at time t:

$$p = (p_1, p_2, \dots, p_K), \quad p_i \geq 0, \quad \sum_{i=1}^{K} p_i = 1, \tag{5.12}$$

where K denotes the number of possible score levels. Let $i^* = \arg\max_i p_i$ be the index of the most likely state (the mode). Variance inflation introduces a local Gaussian-like kernel centred around this mode, which assigns higher weights to states near the mode and lower weights to those further away:

$$k_i = \exp\left(-\frac{(i - i^*)^2}{2}\right), \quad i = 1, \ldots, K. \tag{5.13}$$

The kernel is normalised to ensure it defines a valid probability distribution:

$$k_i = \frac{k_i}{\sum_{j=1}^{K} k_j}. \tag{5.14}$$

The final variance-inflated distribution is obtained by blending the original probabilities with the kernel according to a mixing parameter $\gamma \in [0, 1]$:

$$\tilde{p}_i = \gamma p_i + (1 - \gamma) k_i, \quad i = 1, \ldots, K. \tag{5.15}$$

The parameter γ controls the degree of variance inflation:

- When $\gamma = 1$, the CPT remains unchanged, and the original distribution is preserved.
- When $\gamma = 0$, the CPT is replaced entirely by the symmetric kernel, which spreads probability mass around the mode in a unimodal, locally concentrated fashion.
- For typical values $\gamma \in [0.7, 0.9]$, the CPT retains its original structure but is softened, with peak probabilities reduced and neighbouring states elevated.

Figure 5.5 compares the original probability distribution with the altered one when parameter $\gamma = 0.7$.

Variance inflation also complements other transformations such as skewing or temperature scaling. For example, skewing biases the distributions toward deterioration, while variance inflation ensures that this bias is not expressed as a deterministic one-way trajectory but retains some degree of local randomness.

5.1.6 Scenario 6: Structural Modifications

The final non-stationary scenario departs from adjustments to the CPTs and instead introduces changes directly at the structural level of the DBN graph. While the earlier scenarios maintained a fixed network topology and altered

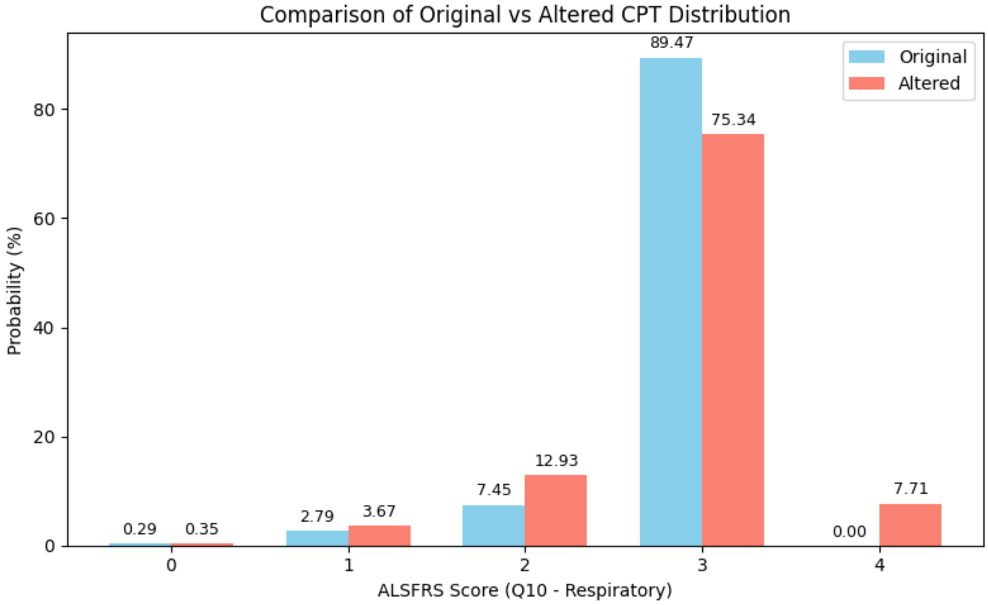

Figure 5.5: Comparison of the original and altered CPT distributions for `Q10 - Respiratory`, using Variance Inflation with $\gamma = 0.7$.

only the numerical values of the conditional probability distributions, here the parent-child relationships between variables themselves evolve across disease phases. This simulates a particularly challenging form of concept drift, where not only the likelihood of outcomes changes over time but the underlying mechanisms that generate these outcomes also shift.

Formally, let the DBN in phase p be represented by a directed acyclic graph $G^{(p)} = (V, E^{(p)})$, where V is the set of variables and $E^{(p)}$ is the set of directed edges in phase p. Non-stationarity of this type implies that:

$$G^{(0)} \neq G^{(1)} \neq G^{(2)}, \tag{5.16}$$

meaning that the dependency structure is phase-specific. This evolution reflects the clinical observation that the drivers of disease progression in ALS may differ at various stages of the disease.

To model such causal shifts, each phase is associated with its own set of mandatory edge constraints. Let $M^{(p)} \in \{0,1\}^{n \times n}$ denote the mandatory adjacency matrix for phase p, where:

$$M_{i,j}^{(p)} = \begin{cases} 1 & \text{if edge } X_i \to X_j \text{ must exist in phase } p, \\ 0 & \text{otherwise.} \end{cases} \tag{5.17}$$

The structure for each phase is then learned using score-based search with the BIC as the objective function, subject to the inclusion of mandatory edges:

$$G^{(p)} = \arg\max_{G} \mathrm{BIC}(G, D) \quad \text{such that } M^{(p)} \subseteq G. \tag{5.18}$$

The outcome of this procedure is therefore a set of phase-specific DBNs $\{G^{(0)}, G^{(1)}, G^{(2)}\}$, each with its own topology and CPTs. These graphs simulate a world where the underlying causal mechanisms of ALS are not constant but instead evolve as the disease progresses.

5.2 Relevance of the Non-Stationary Scenarios to Disease Modelling

The six non-stationary scenarios introduced in Section 5.1 are designed not only as theoretical perturbations of the DBN but also as clinically meaningful stress tests of predictive modelling approaches. Each scenario is associated with a distinct clinical interpretation of ALS progression, while simultaneously probing specific weaknesses or robustness properties of deep sequence models. Taken together, they provide a systematic framework for evaluating how well models can generalise under uncertainty, variability, and structural change.

CPT flattening introduces a controlled increase in entropy in the DBN, effectively simulating diverse patterns of progression. From an experimental perspective, this scenario provides a principled mechanism to evaluate how well sequential deep learning models adapt to changes in the underlying probability structure. Specifically, flattening tests whether models can preserve predictive accuracy when the signal-to-noise ratio decreases and outcomes become less deterministic. Furthermore, it probes model robustness to population heterogeneity, as the simulated patients generated under this transformation no longer follow tightly constrained progression patterns.

By strongly reweighting probabilities through the parameter γ, the extreme skewing scenario generates trajectories with accelerating, nonlinear decline. This scenarios is useful to stress-test predictive models under abrupt and asymmetric changes in the progression dynamics. From a clinical standpoint, this scenario corresponds to situations where patients initially experience a slow decline, followed by a sudden and steep deterioration in specific functional domains. Such sharp, nonlinear deteriorations can represent one of the greatest challenges for clinicians, as they require rapid adjustments in care and interventions. The probabilistic transitions shift from stationarity, such that the past trajectory is no longer a reliable indicator of the future. For sequential models, this poses a distributional shift problem - patterns learned from earlier visits may be insufficient to anticipate the abrupt acceleration in decline.

Temperature scaling manipulates the entropy of CPTs directly, creating either sharply deterministic or highly uncertain distributions. Clinically, this mirrors variability between patients with stable, predictable decline versus those with noisy, unstable trajectories. From an experimental perspective, the goal of this scenario is twofold. First, it enables an evaluation of whether sequence models can maintain predictive accuracy across shifts in entropy, thereby revealing their sensitivity to the confidence structure of the data. Second, it provides insight into how these models handle the transition between noisy, uncertain phases and sharply deterministic phases, a challenge that closely mirrors the real-world variability observed in ALS patients. By systematically varying the temperature parameter across phases, the experiment establishes controlled conditions under which model robustness to uncertainty and determinism can be rigorously assessed.

Step-size banding imposes hard limits on how much functional decline can occur between consecutive visits. This encodes the clinical principle of gradual deterioration, where large jumps in function are implausible. From an experimental perspective, step-size banding introduces a unique type of non-stationarity distinct from the probability reweighting of flattening, skewing, or temperature scaling. Here, the challenge lies not in distributional shifts across outcome space but in structural constraints on transition feasibility. Predictive models must learn to adapt to an environment in which long-range jumps are impossible, meaning that effective prediction requires a fine-grained understanding of local temporal dynamics. The overarching goal of this scenario is therefore to test the capacity of sequence models to forecast under conditions where gradual deterioration is enforced and large jumps are prohibited.

Variance inflation introduces small stochastic deviations around expected trajectories, producing fluctuations without permitting implausible long-range jumps. Clinically, this reflects the heteroscedastic nature of ALS progression, where some patients deviate unpredictably from expected courses due to complications or treatment responses. Models trained under this scenario must demonstrate robustness to local variability and uncertainty. Furthermore, models that can handle variance-inflated trajectories are more likely to generalise to real-world clinical settings, where uncertainty and local deviations from expected progression are unavoidable.

The last scenario enforces phase-dependent changes to the DBN structure itself. Clinically, this scenario mirrors the fact that ALS progression is not driven by a single static mechanism but by a complex interplay of factors that emerge or recede in importance across phases. Structural shifts in the DBN therefore encode a richer, more dynamic view of ALS as a multi-phase disease with evolving causal dependencies. The experimental objective of this scenario is to evaluate whether deep sequence models can handle non-stationarity of causal structure, in addition to distributional shifts. Successfully predicting in this context requires not only temporal modelling capacity but also adaptability to changing interdependencies between variables. This scenario thus represents the most difficult stress-test in the experimental suite, probing whether models can generalise across phases where the "rules of the game" themselves change.

Table 5.1: Summary of the six non-stationary scenarios: clinical interpretation and experimental objective.

Scenario	**Clinical Meaning**	**Experimental Objective**
CPT Flattening	Later phases less predictable, reflecting heterogeneous progression	Test robustness to reduced signal-to-noise ratio and outcome variability
Extreme Skewing	Accelerating decline after a stable period	Evaluate adaptability to abrupt distributional shifts and nonlinear deterioration
Temperature Scaling	Stable vs. noisy patients with variable determinism	Probe sensitivity to entropy regimes and robustness across uncertain vs. peaked distributions
Step-Size Banding	Gradual functional decline, no implausible large jumps	Assess ability to learn fine-grained local dynamics under structural constraints
Variance Inflation	Local deviations from expected trajectory, late-phase heterogeneity	Test robustness under heteroscedasticity and fluctuating progression patterns
Structural Modifications	Different clinical drivers across phases	Evaluate ability to adapt to evolving causal structure beyond distributional shifts

Table 5.1 summarises the six scenarios, outlining their clinical interpretation and the corresponding experimental objectives.

5.3 Simulation of Synthetic Patient Trajectories

Once the baseline DBN and its scenario-specific variants were created, the next step was to generate synthetic patient trajectories suitable for training and evaluating predictive models. The objective was to create datasets that mirror the temporal structure of longitudinal ALS trials, both in terms of visit scheduling and phase transitions, while also incorporating controlled sources of non-stationarity.

Each scenario was parameterised in a way that reflected phase-dependent differences in clinical progression, thereby embedding non-stationarity directly into the generative process. In single-phase experiments, patients remained in a single disease stage, and non-stationarity was introduced by altering parameters between the first and second halves of their follow-up visits. In contrast, multi-phase and controlled cohort experiments simulated phase-dependent progression, with parameters adapted across Phase 0 (early), Phase 1 (intermediate), and Phase 2 (advanced).

Tables 5.2 and 5.3 summarise the parameterisation used in each scenario for single-phase and multi-phase/controlled cohort datasets, respectively.

Table 5.2: Parameter modifications for each single-phase dataset simulation scenario. Non-stationarity is introduced by applying baseline parameters in the first half of the visits and altered parameters in the second half.

Scenario	**CPT Transformation**	**First half of visits**	**Second half of visits**
1	CPT Flattening	$\alpha = 1$	$\alpha = 0.8$
2	Extreme Skewing	$\gamma = 0$	$\gamma = 2$
3	Temperature Scaling	$t = 1$	$t = 1.2$
4	Step-size Banding	$J = 1$, $\epsilon = 0.05$	$J = 2$, $\epsilon = 0.05$
5	Variance Inflation	$\gamma = 1$	$\gamma = 0.7$
6	Structural Modifications	TSO $\rightarrow$ Q1-Q10	TSO, Age, Onset Site $\rightarrow$ Q1-Q10

The choice of parameter values in each scenario was informed by both clinical intuition and the need to stress-test models systematically:

Table 5.3: Parameter modifications for each multi-phase and controlled cohort dataset simulation scenario. Patients transition between Phase 0, Phase 1, and Phase 2, with parameters adapted to reflect stage-specific progression dynamics.

Scenario	CPT Transformation	Phase 0	Phase 1	Phase 2
1	CPT Flattening	$\alpha = 1$	$\alpha = 0.9$	$\alpha = 0.8$
2	Extreme Skewing	$\gamma = 1$	$\gamma = 2$	$\gamma = 4$
3	Temperature Scaling	$t = 1$	$t = 1.2$	$t = 1.4$
4	Step-size Banding	$J = 0, \epsilon = 0.05$	$J = 1, \epsilon = 0.05$	$J = 2, \epsilon = 0.05$
5	Variance Inflation	$\gamma = 1$	$\gamma = 0.8$	$\gamma = 0.6$
6	Structural Modifications	Onset site $\rightarrow$ Q1-Q10	Age $\rightarrow$ Q1-Q10	Sex $\rightarrow$ Q1-Q10

- **CPT Flattening:** α values decrease from early ($\alpha = 1$) to late disease ($\alpha = 0.8$), reflecting that early phases are stable whereas late phases are heterogeneous and uncertain.
- **Extreme Skewing:** γ increases from either $\gamma = 1$ in early phase to $\gamma = 4$ in last phase, or from $\gamma = 0$ to $\gamma = 2$ (depending on the dataset), mimicking accelerating decline in later stages.
- **Temperature Scaling:** Temperature rises from $t = 1$ in the beginning to $t = 1.2$ and $t = 1.4$ in later phases, introducing higher entropy and noisier dynamics in later disease.
- **Step-Size Banding:** The step-size J increases across phases from $J = 0$ or $J = 1$ to $J = 2$ in the last phase, encoding the principle that decline is gradual but becomes slightly more variable in advanced disease.
- **Variance Inflation:** Variance inflation reduces γ from 1 in the initial phase to 0.7 and 0.6 in the last phases of the disease progression, embedding greater late-phase variability around expected decline.
- **Structural Modifications:** Mandatory edges were imposed in a phase-specific way: onset site $\rightarrow$ Q1-Q10 in Phase 0, age $\rightarrow$ Q1-Q10 in Phase 1, and sex $\rightarrow$ Q1-Q10 in Phase 2, reflecting shifting clinical drivers. This

represents the clinical intuition that whether ALS begins in bulbar or limb regions strongly shapes early functional decline. In mid-disease phase, the trajectory of functional deterioration is modulated by the patient's age, while in late-phase decline may differ by sex.

The simulation process involved repeatedly sampling the next visit state of each patient from the CPTs of the network, thereby unrolling longitudinal sequences of varying lengths. The process began with sampling an initial observation from the joint distribution encoded in the DBN, which included static baseline variables and the dynamic variables for the first two visits t_1 and t_2. From the third visit onwards, each subsequent state was generated by:

1. Selecting the relevant network (baseline or scenario-altered) depending on the experimental condition.
2. Shifting dynamic variables from time step t_i to time step t_{i-1}.
3. Computing the conditional distribution of each dynamic variable at t_i given its parents values at t_{i-1}.
4. Sampling a new state from the corresponding CPT.

This recursive mechanism yielded full longitudinal trajectories in which every visit included both static information and the simulated dynamic variables that evolved over time.

For every scenario, two parallel dataset versions were created:

- **Full-alteration version:** CPTs for all ALSFRS items (Q1-Q10) were modified in the alternative DBNs.
- **Targeted-alteration version:** only CPTs and parent sets of Q1 (Speech), Q2 (Salivation), Q3 (Swallowing), and Q10 (Respiratory) were modified.

This dual setup allowed assessment of predictive models both under global non-stationarity and under localised perturbations affecting only selected functional domains.

Each dataset consisted of between 1000 and 4800 simulated patients, depending on whether single-phase or multi-phase designs were employed. Reproducibility was enforced by fixing the random seed prior to sampling.

In the simplest case, patients were assumed to remain within a single disease phase for their entire follow-up. First, a baseline dataset with 1000 patients was simulated directly from the unaltered DBN. Then, for each probability perturbation scenario, 1000 patients were generated, forming six new datasets. To simulate realistic heterogeneity in trial participation, the number of visits per patient was drawn from a truncated normal distribution:

$$N \sim \max(5, \mathcal{N}(\mu = 10, \sigma = 2)), \tag{5.19}$$

ensuring that every patient had at least five visits. In practice, this produced trajectories of variable length, with the majority of patients spanning between 8 and 12 visits, consistent with typical clinical trial follow-up windows in the PRO-ACT dataset. In every scenario, the baseline network was used during the first half of the visits, and the altered network during the second half, thereby introducing non-stationarity mid-sequence.

To capture the progressive nature of ALS, a more flexible framework was developed to simulate patients transitioning between predefined disease phases. Each phase path was defined as an ordered sequence of phase labels (e.g., $(0, 1)$ or $(0, 1, 2)$), with the number of visits per phase again sampled from a truncated normal distribution $\mathcal{N}(10, 2)$ with a minimum of five visits. A total of 4800 patients were generated according to the following case distribution:

- 1000 patients remaining in Phase 0 only,
- 1000 patients remaining in Phase 1 only,
- 250 patients remaining in Phase 2 only,
- 2000 patients transitioning from Phase 0 to 1 ($0 \to 1$),
- 500 patients transitioning from Phase 1 to 2 ($1 \to 2$),
- 50 patients transitioning from Phase 0 to 1 and then to 2 ($0 \to 1 \to 2$).

The *Time Since Onset* (TSO) variable was updated accordingly to encode the current phase, and the appropriate phase-specific network was selected.

Finally, additional datasets, called controlled cohort datasets, were constructed with fixed allocations to enable controlled comparisons. In this case, 3000 patients were generated along the full disease trajectory ($0 \to 1 \to 2$) with exactly 7 visits per phase. This yielded uniform 21-visit trajectories, providing a benchmark setting without variability in follow-up intensity.

Patients were then stratified by their disease phase or phase-path (*TSO*) and split into *training* (60%), *validation* (20%), and *test* (20%) sets. The stratification ensured a balanced representation of phases across splits, which was crucial for enabling robust phase-specific evaluation.

6

Predictive Modelling of Disease Progression

6.1 Single-Layer LSTM and GRU Models

Recurrent neural networks (RNNs) represent one of the most fundamental classes of deep sequence models. They are specifically designed to handle sequential data, where observations are not independent but ordered in time. In the context of disease progression modelling, this temporal dependency is crucial because the clinical state of a patient at a given visit is not random but conditioned on the history of prior visits [12]. However, standard RNNs are known to suffer from severe limitations, most notably the vanishing and exploding gradient problem during backpropagation through time, which hinders their ability to learn long-term dependencies [8]. Gradients tend to decay (or blow up) exponentially with sequence length, causing the model to focus primarily on recent inputs and lose information from distant time steps. To overcome these issues, more advanced recurrent architectures have been developed, among which the Long Short-Term Memory (LSTM) and the Gated Recurrent Unit (GRU) are the most widely used [28, 17]. Both models introduce gating mechanisms that regulate the flow of information through the recurrent cell, enabling them to selectively retain, update, or discard information across time steps.

6.1.1 Long Short-Term Memory (LSTM)

The LSTM architecture extends the standard RNN by introducing an explicit cell state that acts as a memory buffer, together with three multiplicative gates that control the flow of information [28]:

- The *forget gate* f_t determines how much of the previous cell state c_{t-1} should be retained or discarded.
- The *input gate* i_t decides which components of the new candidate state $\tilde{c}_t$ should be incorporated.
- The *output gate* o_t regulates how much of the updated cell state c_t is exposed as the hidden state h_t.

Formally, given an input vector x_t at time step t and the previous hidden and cell states (h_{t-1}, c_{t-1}), the LSTM updates are defined as:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f), \tag{6.1}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), \tag{6.2}$$

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c), \tag{6.3}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \tag{6.4}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o), \tag{6.5}$$

$$h_t = o_t \odot \tanh(c_t). \tag{6.6}$$

Here, $\sigma(\cdot)$ denotes the logistic sigmoid function, $\tanh(\cdot)$ the hyperbolic tangent, and $\odot$ the element-wise product. This gating structure allows the LSTM to regulate the degree to which past information influences the current predictions, enabling it to capture long-term dependencies in clinical trajectories such as gradual decline across several visits. The architecture of the LSTM cell is illustrated in Figure 6.1 [21].

6.1.2 Gated Recurrent Unit (GRU)

The GRU can be seen as a streamlined variant of the LSTM [17] that reduces complexity while retaining the essential gating functionality. Instead of maintaining separate cell and hidden states, the GRU merges them into a single hidden state h_t. The architecture introduces two gates:

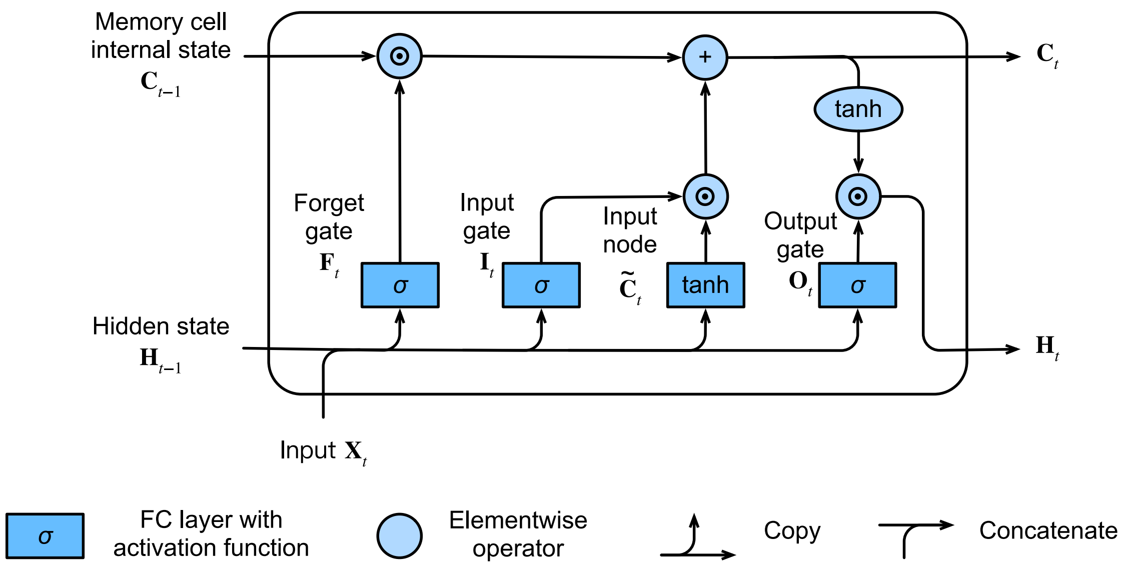

Figure 6.1: Architecture of the LSTM cell.

- The *update gate* z_t controls how much of the past information should be carried forward.

- The *reset gate* r_t determines how much of the past should be ignored when computing the candidate hidden state.

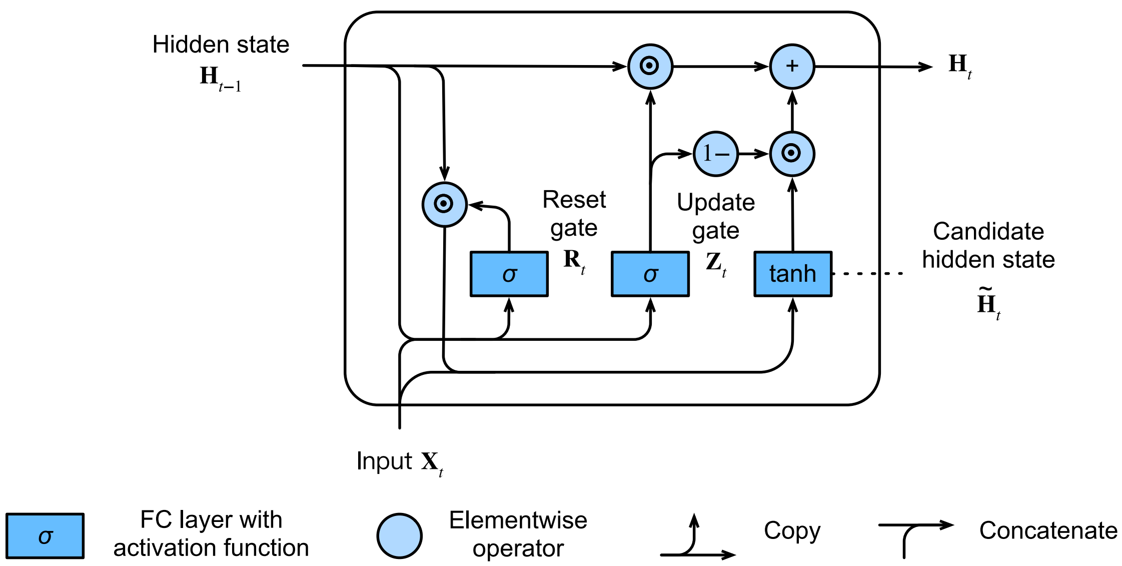

Figure 6.2: Architecture of the GRU cell.

Mathematically, the GRU updates are expressed as:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z), \tag{6.7}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r), \tag{6.8}$$

$$\tilde{h}_t = \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h), \tag{6.9}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t. \tag{6.10}$$

The update gate z_t serves a role analogous to the LSTMs input and forget gates combined, while the reset gate r_t allows the network to selectively ignore past states when necessary. This more compact formulation makes GRUs computationally more efficient than LSTMs while maintaining strong performance on many sequential modelling tasks. The architecture is shown in Figure 6.2 [21].

6.1.3 Handling of Padded Sequences

One of the challenges in applying recurrent models in clinical data lies in the variable sequence lengths across patients [27]. In real-world or simulated longitudinal datasets, some patients contribute only a few visits, whereas others are followed over many time points. To enable batch processing, all sequences must be aligned to a common length, which is typically achieved by padding shorter sequences with zeros. However, naively feeding padded sequences into the network would mislead the recurrent dynamics, as the model would interpret the padding as actual observations.

To address this, a masking mechanism is applied, where the padding values are explicitly ignored during training and evaluation. This allows the model to focus exclusively on the meaningful parts of each sequence, preserving the integrity of patient trajectories while enabling efficient training across heterogeneous cohorts.

6.2 Stacked LSTM and GRU Models

While single-layer RNNs such as LSTMs and GRUs provide a robust framework for modelling temporal dependencies, they are often limited in their ability to capture the full complexity of long, heterogeneous disease trajectories. Clinical progression is rarely driven by a single dominant temporal pattern. Rather, it reflects a hierarchy of processes that unfold across different timescales, ranging from short-term fluctuations in functional ability to long-term gradual decline. To address this, researchers have extended recurrent architectures by stacking multiple layers of LSTM and GRU units, thereby increasing the temporal depth and representational capacity of the model [7].

6.2.1 Multi-layer Recurrence and Temporal Depth

In a stacked recurrent architecture, the hidden states produced by one recurrent layer are passed as inputs to the next layer, creating a hierarchical processing pipeline. Formally, given a sequence of inputs $\{x_t\}_{t=1}^{T}$, the hidden state at layer l and time step t is computed as:

$$h_t^{(l)} = f^{(l)}(h_{t-1}^{(l)}, h_t^{(l-1)}), \tag{6.11}$$

where $f^{(l)}$ denotes the recurrent update function (either LSTM or GRU) for the l-th layer, and $h_t^{(0)} = x_t$. By stacking L such layers, the model can extract increasingly abstract temporal features, with lower layers capturing short-term dependencies (e.g., transitions between consecutive visits) and higher layers modelling long-term dependencies (e.g., overall disease trajectory). The architecture is illustrated in Figure 6.3 [21].

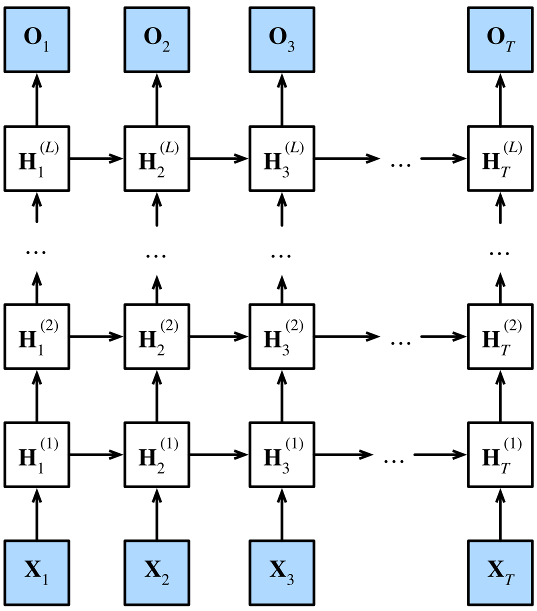

Figure 6.3: Architecture of a deep RNN.

6.2.2 Layer Normalisation and Regularisation

Stacked architectures, however, introduce new challenges, particularly in terms of stability during training. As signals are propagated through multiple recurrent layers, they can become distorted, leading to difficulties in optimisation and susceptibility to overfitting. To mitigate these issues, *layer normalisation* is commonly applied within each recurrent layer.

Layer normalisation [5] standardises the summed inputs to a recurrent unit across the hidden dimension, ensuring that the scale and distribution of activations remain stable regardless of sequence length or batch composition. Given a hidden state vector $h \in \mathbb{R}^d$, layer normalisation rescales it as:

$$\hat{h}_i = \frac{h_i - \mu}{\sqrt{\sigma^2 + \epsilon}}, \tag{6.12}$$

$$h_i^{\text{norm}} = \gamma \hat{h}_i + \beta, \tag{6.13}$$

where μ and σ^2 are the mean and variance across the hidden dimension, ϵ is a small constant for numerical stability, and γ, β are trainable parameters. This normalisation not only improves convergence speed but also reduces internal covariate shift, enabling the model to generalise more reliably across heterogeneous patient trajectories.

In addition to normalisation, dropout regularisation is often applied between layers to reduce overfitting [49]. This is particularly important in clinical modelling, where the number of patients may be relatively limited compared to the number of parameters in deep recurrent networks.

6.3 Recurrent Models with Attention

While stacked RNNs enhance the representational power of sequential models, they still suffer from a fundamental limitation - their reliance on sequential propagation of hidden states. Even in multi-layer configurations, information must pass step by step through time, which can cause long-range dependencies to be diluted or forgotten. This constraint is particularly problematic in the context of disease progression, where the deterioration of specific functional items may depend not only on the immediately preceding state but also on distant historical patterns. To address these limitations, recurrent models can be augmented with self-attention mechanisms, allowing them to selectively access and reweight information from across the entire sequence without relying solely on recurrent propagation.

6.3.1 Integration of Self-Attention into Recurrent Architectures

The central idea of self-attention is to compute, for each time step, a weighted combination of hidden representations from all other time steps in the sequence. This mechanism provides the model with the ability to capture global dependencies that may be difficult for recurrence alone to preserve [55, 6]. Formally, given a hidden sequence $H = \{h_t\}_{t=1}^{T}$ produced by an LSTM or GRU layer, the self-attention mechanism constructs three projections: queries $Q = HW^Q$, keys $K = HW^K$, and values $V = HW^V$, where W^Q, W^K, W^V are trainable weight matrices. The attention weights are computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V, \tag{6.14}$$

where d_k is the dimensionality of the key vectors, ensuring proper scaling of dot products. The resulting representation for each time step is a weighted sum of values from all positions in the sequence, with the weights reflecting their contextual relevance. The architecture of scaled dot-product and multi-head attention is illustrated in Figure 6.4 [55].

When integrated with recurrent models, this mechanism operates on the hidden states already enriched with temporal information through recurrence, thus combining local temporal encoding (captured by RNNs) with global relational reasoning (enabled by attention). This hybrid design is particularly advantageous in ALS progression modelling since it allows the model to simultaneously track fine-grained local changes in functional items and long-range dependencies between functional domains across visits.

6.3.2 Handling Clinical Sequences: Causal and Padding Masks

A key consideration when applying attention to clinical data is the presence of irregularly spaced visits and variable sequence lengths across patients. Directly applying self-attention without constraints could inadvertently allow

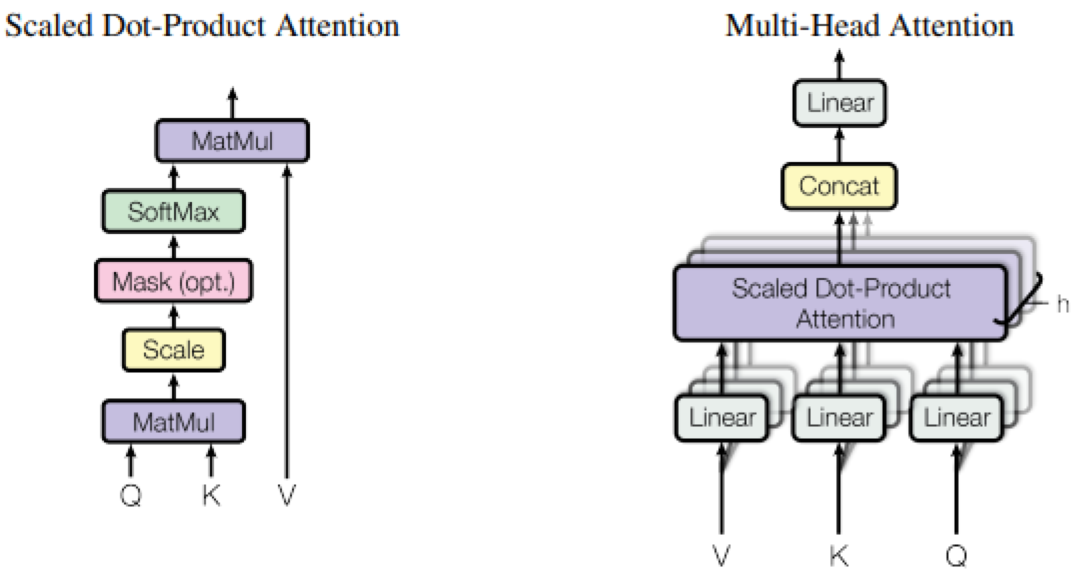

Figure 6.4: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

information from the future to influence predictions of the past, violating the causal structure of disease progression. To preserve temporal consistency, a *causal mask* is applied, ensuring that at time step t, the model only attends to hidden states from time steps $\leq t$. This constraint is implemented by assigning negative infinity to attention logits corresponding to future positions, effectively preventing the model from "peeking ahead" in the sequence.

In addition to causality, clinical datasets often require handling padded sequences, since patients differ in the number of visits they undergo. To ensure that padded positions do not interfere with model computations, a *padding mask* is used. This mask assigns zero weight to padded entries, preventing them from contributing to the attention context or from influencing the hidden state dynamics.

In practice, the two masks (causal and padding) are combined to form a joint masking strategy that simultaneously enforces temporal consistency and robustness to variable sequence lengths. Mathematically, for an attention logit matrix $A \in \mathbb{R}^{T\times T}$, the masked version is:

$$A_{ij} = \begin{cases} -\infty & \text{if } j > i \quad \text{(causality)} \\ -\infty & \text{if position } j \text{ is padding} \\ A_{ij} & \text{otherwise,} \end{cases} \tag{6.15}$$

ensuring that the final attention distribution is normalised only over valid historical positions.

6.4 Transformer Models

The limitations of recurrent models, even when enhanced with self-attention mechanisms, become evident when scaling to longer sequences or attempting to capture highly non-local dependencies. Recurrent architectures process inputs sequentially, propagating information step by step through hidden states, which inherently restricts parallelism and makes learning long-term dependencies vulnerable to vanishing gradients. In contrast, the Transformer architecture represents a paradigm shift in sequence modelling, dispensing entirely with recurrence and convolution, and instead relying exclusively on stacked layers of self-attention and feed-forward transformations [55]. This design has proven remarkably effective in natural language processing and, more recently, in healthcare time-series modelling, where irregular, high-dimensional, and long-range dependencies are prevalent [35].

At its foundation, the Transformer operates on a sequence of input embeddings $x_1, x_2, \ldots, x_T$, each representing the state of a patient at visit t. Since Transformers do not have an inherent notion of order, positional encodings are added to the embeddings to preserve temporal structure. In the original formulation, sinusoidal encodings were employed:

$$PE_{(t,2i)} = \sin\left(\frac{t}{10000^{2i/d}}\right), \tag{6.16}$$

$$PE_{(t,2i+1)} = \cos\left(\frac{t}{10000^{2i/d}}\right), \tag{6.17}$$

where d is the embedding dimension and i indexes the embedding components. In modern adaptations, as it is also done in this work, these fixed sinusoidal encodings are often replaced with learnable positional embeddings, allowing the model to directly optimise temporal representation from data. Formally, let $x_t \in \mathbb{R}^d$ denote the projected feature vector at timestep t. The learnable positional encoding adds an embedding vector p_t to each input:

$$\tilde{x}_t = x_t + p_t. \tag{6.18}$$

Here, p_t is learned jointly with all other model parameters, allowing the network to adaptively weight temporal structure rather than relying on a fixed sinusoidal basis.This flexibility is particularly valuable for clinical data, where visit intervals may not be equidistant and disease progression may not follow uniform patterns.

The core of the Transformer is the *multi-head self-attention* mechanism, which extends the single-head formulation introduced earlier. Instead of computing a single attention distribution, the hidden representations are projected into multiple subspaces, each corresponding to an attention "head":

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O, \tag{6.19}$$

where each head is defined as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V). \tag{6.20}$$

By combining information from multiple attention heads, the model is able to capture diverse types of relationships: some heads may focus on short-range continuity, while others specialise in long-range associations across visits or functional domains. The outputs of attention layers are then processed through position-wise feed-forward networks, normalisation layers, and residual connections, forming the repeating Transformer block.

Mathematically, for each block, the transformation can be summarised as:

$$H' = \text{LayerNorm}\left(H + \text{MultiHead}(H, H, H)\right), \tag{6.21}$$

$$H'' = \text{LayerNorm}\left(H' + \text{FFN}(H')\right), \tag{6.22}$$

where H represents the sequence of embeddings entering the block, and FFN is a two-layer feed-forward network with non-linear activation. Residual connections stabilise training and facilitate gradient flow, while layer normalisation ensures consistent scaling of hidden representations across time steps. The full architecture is shown in Figure 6.5 [55].

6.4.1 Differences Compared to Recurrent Models

The Transformer architecture differs fundamentally from recurrent models in several ways that directly impact its suitability for modelling disease progression:

- **Parallelism:** Unlike RNNs, which process sequences step by step, Transformers operate on all positions simultaneously, enabling efficient parallelisation on modern hardware. This property is particularly advantageous for hyperparameter optimisation across large synthetic datasets, where computational efficiency is critical.
- **Long-Range Dependency Modelling:** Recurrent models rely on hidden state propagation to capture distant dependencies, which can be unstable over long horizons. Transformers, by contrast, compute pairwise attention between all time steps directly, allowing them to preserve relationships across arbitrarily long sequences without degradation. This is crucial for modelling of disease trajectories, where early clinical signs may influence outcomes observed many visits later.
- **Flexibility in Handling Non-Uniform Sequences:** By decoupling positional information from the model structure and instead encoding it explicitly, Transformers can more easily adapt to irregular visit schedules or phase-dependent changes in disease dynamics.
- **Absence of Sequential Bias:** Whereas RNNs inherently emphasise sequential continuity, Transformers treat all positions symmetrically, with ordering provided solely by positional encoding. This design allows for more flexible modelling of non-stationary trajectories, where dependencies may abruptly shift across disease phases.

6.5 Relevance to Disease Modelling

The different deep sequence architectures described in this chapter can be positioned along a continuum of increasing expressiveness and flexibility for modelling disease progression. Their relative strengths and limitations become

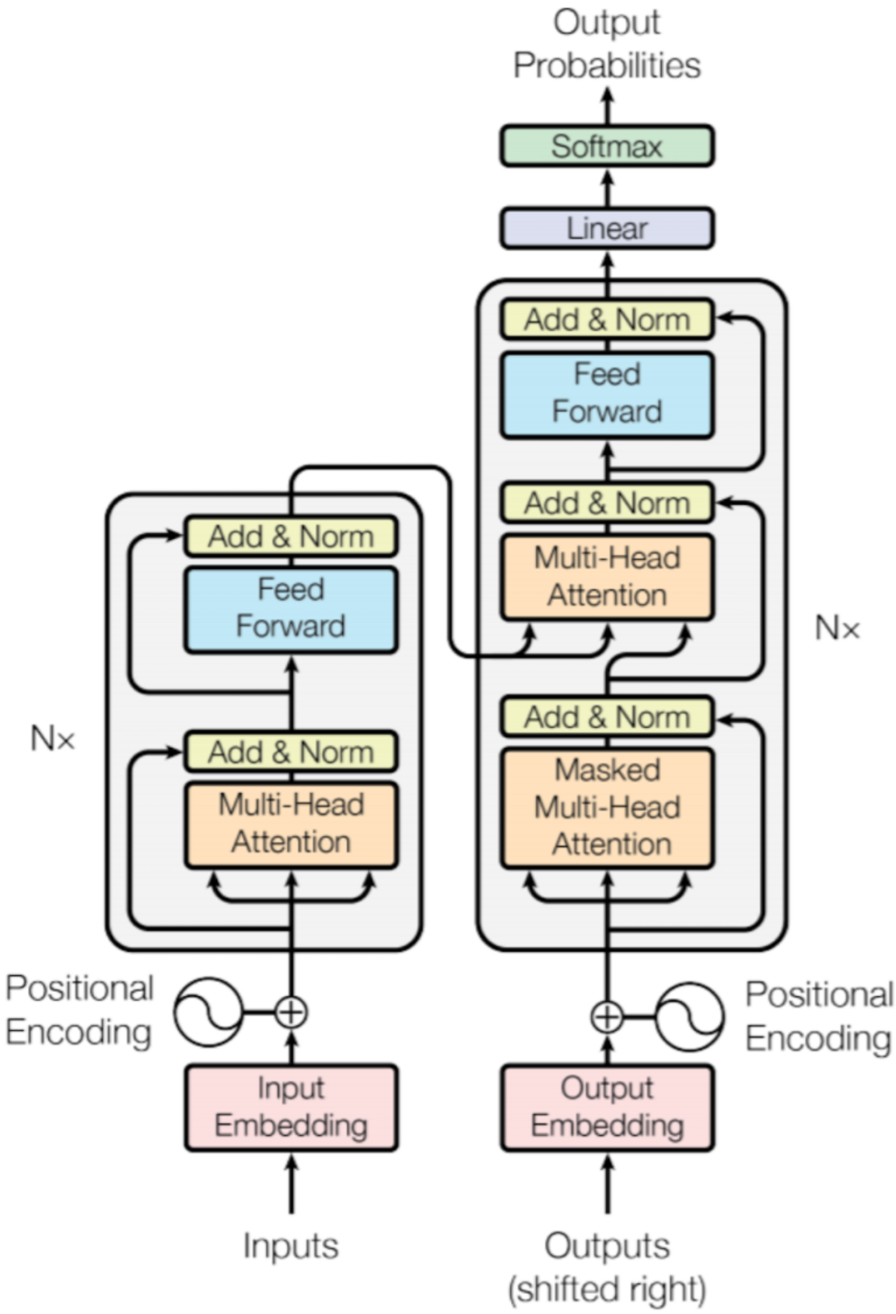

Figure 6.5: The Transformer model architecture.

particularly evident when placed in the context of clinical requirements, such as handling irregular follow-up schedules, capturing heterogeneous progression patterns, and adapting to non-stationarity across disease phases.

The choice of single-layer LSTM and GRU models provides a natural baseline for investigating the predictability of disease progression. Their relatively simple structure, combined with powerful gating mechanisms, makes them well-suited to capturing dominant temporal dependencies without overfitting. Furthermore, their reliance on masking ensures robustness when dealing with irregular follow-up schedules, which are common in clinical datasets. By first evaluating single-layer recurrent models, we establish a benchmark against which more complex architectures can be systematically compared.

Building on this foundation, stacked LSTM and GRU models offer several advantages. By leveraging multi-layer recurrence, the network can encode subtle interactions between functional domains that evolve over different timescales. This allows the models to capture both gradual long-term decline and short-term fluctuations in disease progression. Layer normalisation further ensures stability in learning, enabling effective training even in the presence of patient heterogeneity and irregular visit schedules. From a clinical perspective, stacked recurrent models thus should provide a more faithful representation of the disease process and form a natural next step beyond single-layer models.

The integration of self-attention mechanisms into recurrent architectures enhances the ability to model disease progression in ways that align closely with clinical intuition. Attention allows the model to dynamically identify which historical visits or functional items are most informative for predicting the next state, effectively simulating how clinicians often rely on both recent and distant patient history when assessing prognosis. By combining recurrence with attention, these hybrid models retain the strengths of RNNs in modelling local continuity and smooth transitions, while gaining the flexibility of attention to capture long-range, non-local dependencies. This makes the approach especially powerful for disease progression modelling under non-stationary conditions, where patterns of dependence may shift across phases and patient outcomes may be influenced by both recent and distant events.

Finally, the Transformer provides a fundamentally different approach to temporal modelling, dispensing with recurrence altogether. Its ability to directly reweight the importance of past visits without sequential bottlenecks means it can adapt more effectively to sudden changes in conditional distributions or causal structures. For instance, when disease enters a rapid progression phase, the Transformer can shift attention away from distant stable visits and focus on the most recent signs of decline. Conversely, in phases where long-term history is predictive, it can integrate evidence across the entire sequence without loss of information. In this sense, the Transformer offers a particularly strong testbed for robustness under non-stationarity, trading sequential inductive biases for global context modelling and parallel computation.

In summary, the architectures considered in this work can be viewed as complementary steps toward increasingly powerful representations of disease progression. Table 6.1 summarises the main strengths and limitations of each architecture in the context of disease progression modelling.

Table 6.1: Comparison of deep sequence architectures for disease progression modelling.

Model	**Strengths**	**Limitations**
Single-layer LSTM / GRU	Robust and interpretable baseline	Limited long-range modelling
Stacked LSTM / GRU	Hierarchical and multi-timescale dynamics	Risk of overfitting, training instability
RNNs with Attention	Combines local + global dependencies	Increased computational cost
Transformer	Parallelism, long-range modelling, non-stationary adaptability	Requires larger datasets

6.6 Hyperparameter Optimisation

The predictive performance of deep learning models depends not only on the network architecture but also on the careful selection of hyperparameters that govern model capacity, regularisation, and training dynamics. In the context of recurrent and attention-based models for disease progression forecasting, hyperparameter optimisation is especially critical, since both under-parameterisation (leading to underfitting) and over-parameterisation (leading to overfitting and poor generalisation) are plausible risks when working with longitudinal data.

Rather than manually selecting hyperparameters or relying on exhaustive grid search, this work employs *Bayesian optimisation*, which provides a more principled and efficient exploration of the hyperparameter space. Bayesian optimisation constructs a probabilistic surrogate model of the objective function, in this case the validation loss of the network, and iteratively refines this surrogate as new evaluations are performed [48].

Formally, let $f(\theta)$ denote the objective function that maps a vector of hyperparameters θ to the validation loss. Since direct evaluation of f is expensive, a surrogate model $\hat{f}$, often based on Gaussian processes, is fitted to approximate the true objective landscape. At each iteration, the optimiser selects the next candidate hyperparameters by maximising an acquisition function $a(\theta \mid \hat{f})$, which balances: **exploitation** (focusing on regions where low loss is predicted) and **exploration** (evaluating uncertain regions of the space).

Each trial consists of training the candidate architecture for a fixed number of epochs set to 50 with early stopping using patience of 5 epochs. The optimiser updates its surrogate and acquisition model after each trial, gradually converging toward hyperparameter regions with the most promising performance. This allows the search to be both data-efficient and computationally tractable, even in moderately high-dimensional hyperparameter spaces.

The search space was defined to capture essential degrees of freedom in model design across recurrent, attention-augmented recurrent, and Transformer architectures.

Recurrent Layer Parameters

- Number of recurrent units (hidden size) was explored over powers of two: $\{16, 32, 64, 128\}$.

- Recurrent dropout rates were varied in [0, 0.2], serving as a regularisation mechanism to reduce reliance on any particular temporal dependency.

Dense Layer Parameters

- The number of dense layers stacked after the recurrent backbone was chosen from: $\{1, 2, 3\}$.

- The initial dense width was sampled from $\{32, 64, 128\}$, with subsequent layers halving in size to progressively compress the temporal representations into the prediction space.

- Dropout between dense layers was explored with rates: [0, 0.2].

Attention Parameters

- The number of attention heads was tuned in $\{2, 4, 8, 16\}$, directly controlling the model's capacity to capture heterogeneous dependency structures.

- For Transformers, the embedding dimension was varied over {32, 64, 128}, with the constraint to be divisible by the number of heads.
- Feed-forward network width was searched over: {64, 128}.
- The number of Transformer blocks: {2, 3}, balancing representational depth with training efficiency.
- Global dropout rate: [0, 0.2], applied within both self-attention and feed-forward layers to mitigate overfitting.

Table 6.2: Explored hyperparameter space for recurrent, attention-augmented recurrent, and Transformer architectures.

Layer / Component	**Hyperparameter**	**Search Space / Values**
Recurrent Layer	Number of units (hidden size)	16, 32, 64, 128
	Recurrent dropout	0.0, 0.2
Dense Layer	Number of dense layers	1, 2, 3
	Initial dense width	32, 64, 128
	Dropout between dense layers	0.0, 0.2
Attention / Transformer	Number of attention heads	2, 4, 8, 16
	Embedding dimension	32, 64, 128
	Feed-forward width	64, 128
	Number of Transformer blocks	2, 3
	Global dropout	0.0, 0.2

This systematic hyperparameter search produced optimised configurations tailored to each model family. Importantly, the search space was deliberately structured around clinically meaningful trade-offs: smaller models test robustness under restricted capacity, while larger models probe whether additional expressive power translates into improved prediction of complex, non-stationary trajectories.

Training Procedure

Models were trained using the Adam optimiser [1] and compiled with the Mean Squared Error (MSE) loss, selected due to the continuous and bounded nature of ALSFRS scores. Importantly, the MSE was weighted by timestep masks to ensure that padded values and last-visit targets were excluded from gradient updates. Early stopping with a patience of five epochs was applied to prevent overfitting. For robustness, once the optimal hyperparameter configuration was identified, it was retrained on the combined training and validation sets prior to final evaluation.

6.7 Evaluation Framework

A rigorous evaluation framework was designed to quantify the predictive performance of the models and to capture both global and phase-specific patterns of disease progression. The framework integrates standard regression error metrics, relative error metrics that account for scale, and novel evaluation procedures tailored to the clinical structure of the data, such as phase-specific assessment and transition-aware analysis.

6.7.1 Performance Metrics

To evaluate the accuracy of the predicted ALSFRS items against the true values, several complementary metrics were employed:

- **Mean Squared Error (MSE):** Measures the average squared difference between predicted and true values:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{6.23}$$

  This metric penalises larger errors more heavily, making it sensitive to outliers.

- **Root Mean Squared Error (RMSE):** The square root of MSE, restoring the metric to the same scale as the original items:

$$\text{RMSE} = \sqrt{\text{MSE}} \tag{6.24}$$

- **Mean Absolute Error (MAE):** Represents the average absolute deviation between predictions and true values:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{6.25}$$

Unlike MSE, MAE is less sensitive to outliers and provides a more interpretable notion of typical prediction error.

- **Coefficient of Determination (R^2):** Quantifies the proportion of variance in the true values explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{6.26}$$

Higher values indicate better predictive performance, with $R^2 = 1$ corresponding to perfect predictions.

- **Relative Error Metrics:** To normalise error by the inherent variability in the data, relative versions of the metrics were also computed:

$$\text{rel-MSE} = \frac{\text{MSE}}{\text{Var}(y)}, \quad \text{rel-RMSE} = \frac{\text{RMSE}}{\text{Std}(y)}, \quad \text{rel-MAE} = \frac{\text{MAE}}{\mathbb{E}[|y|]}$$

These metrics facilitate comparability across items and phases with different scales or variances.

6.7.2 Phase-Specific Evaluation

Given the heterogeneity of ALS progression, patients were stratified into disease phases based on *TimeSinceOnset*, with each phase corresponding to a shift in the underlying distribution of ALSFRS scores. In this way, the transition points between two phases are non-stationary points. The evaluation framework computes metrics separately for each phase:

- **Phase 0 (early stage):** Initial visits shortly after onset.
- **Phase 1 (intermediate stage):** Midpoint progression visits.
- **Phase 2 (late stage):** Advanced disease visits.

For each phase, both overall metrics (averaged across items) and per-item metrics were reported. This stratification reveals whether models are more accurate in early disease, where progression is slower, or in later phases, where rapid decline introduces variability.

6.7.3 Transition-Aware Evaluation

In addition to static phase-based evaluation, a transition-aware procedure was introduced to explicitly capture performance around non-stationarity points, i.e., transitions between phases. Specifically, two clinically relevant transitions were considered:

- Phase 0 $\longrightarrow$ Phase 1 (early to intermediate).
- Phase 1 $\longrightarrow$ Phase 2 (intermediate to late).

For each transition, model performance was assessed at relative visit offsets surrounding the transition point:

- $t-2$: two visits before the transition.
- $t-1$: one visit before the transition.
- t: the transition visit.
- $t+1$: one visit after the transition.

This offset-based analysis provides fine-grained insight into the temporal stability of model predictions. It highlights whether models maintain predictive accuracy when disease dynamics change abruptly, which is particularly relevant for evaluating robustness in the presence of non-stationarity.

7 Results

7.1 Overall Model Performance

This section reports the quantitative results obtained for each model architecture on the test set across all simulation scenarios. Results are reported separately for the single-phase dataset, the multi-phase dataset, and the controlled cohort dataset. The analysis focuses on both absolute performance and relative differences across architectures and scenarios. Performance was measured using RMSE, MAE, and R^2. While the Original scenario represents the baseline, the primary interest lies in Scenarios 1-6, which introduce different forms of non-stationarity by perturbing CPTs and structural properties of the DBN used for simulation.

7.1.1 Results on the Single Phase Dataset

Table 7.1 summarises results for the single-phase dataset. Across all models, Scenario 1 (flattening CPTs) produced the most pronounced degradation in performance, with RMSE values increasing substantially (e.g., 0.713 for LSTM and 0.741 for Stacked LSTM) and R^2 dropping below 0.74. Scenario 2 (extreme skewing) was better tolerated, with R^2 values still around 0.87-0.88 across models. Scenario 3 (temperature scaling) also maintained strong performance, with R^2 consistently above 0.89. Scenarios 4 and 5 led to noticeable degradation, particularly in MAE, highlighting that models became less accurate in capturing fine-grained progression dynamics under variance inflation and step-size perturbations.

Notably, Scenario 6 (structural modifications) yielded relatively strong results across models, with performance close to the original baseline (e.g., GRU: RMSE = 0.418, R^2 = 0.913). This suggests that deep sequence models can learn altered dependencies over time.

Table 7.1: Overall Performance on Single Phase Dataset

Architecture	Metric	Original	1	2	3	4	5	6
LSTM	RMSE	0.408	0.713	0.484	0.437	0.499	0.535	0.421
	MAE	0.253	0.459	0.325	0.268	0.292	0.351	0.255
	R^2	0.917	0.736	0.880	0.900	0.870	0.852	0.911
GRU	RMSE	0.411	0.719	0.486	0.440	0.501	0.547	0.418
	MAE	0.251	0.465	0.315	0.275	0.288	0.358	0.255
	R^2	0.916	0.731	0.879	0.899	0.868	0.845	0.913
Stacked LSTM	RMSE	0.426	0.741	0.505	0.457	0.519	0.554	0.437
	MAE	0.272	0.489	0.341	0.293	0.323	0.366	0.277
	R^2	0.910	0.715	0.870	0.891	0.859	0.841	0.905
Stacked GRU	RMSE	0.427	0.732	0.508	0.461	0.519	0.551	0.441
	MAE	0.270	0.478	0.350	0.312	0.309	0.365	0.284
	R^2	0.909	0.722	0.868	0.889	0.859	0.843	0.903
LSTM + Attention	RMSE	0.416	0.716	0.496	0.446	0.506	0.544	0.428
	MAE	0.268	0.451	0.325	0.282	0.296	0.359	0.267
	R^2	0.914	0.734	0.874	0.896	0.866	0.846	0.908
GRU + Attention	RMSE	0.425	0.715	0.502	0.452	0.515	0.548	0.440
	MAE	0.270	0.459	0.344	0.288	0.295	0.359	0.265
	R^2	0.910	0.734	0.871	0.893	0.861	0.844	0.903
Transformer	RMSE	0.419	0.716	0.494	0.453	0.511	0.546	0.439
	MAE	0.258	0.469	0.310	0.293	0.299	0.364	0.270
	R^2	0.913	0.734	0.875	0.893	0.863	0.845	0.904

7.1.2 Results on the Multi Phase Dataset

The multi-phase dataset results are reported in Table 7.2. Models achieved overall lower error compared to the single-phase dataset, with RMSE around 0.38 and R^2 consistently above 0.92 in the original scenario. This suggests that having a larger dataset stabilises learning and improves predictive performance.

Scenario 1 again posed the greatest challenge, but the degradation was less severe than in the single-phase setting, with R^2 values remaining above 0.80. This suggests that additional temporal context from multiple phases provided resilience against flattened CPTs.

Scenario 2 and Scenario 3 showed consistently strong performance ($R^2 > 0.89$ for most models). Scenarios 4 and 5 introduced moderate performance reductions, while Scenario 6 once more demonstrated high robustness, with results nearly indistinguishable from the original baseline. The Transformer achieved the best balance across scenarios, with the lowest RMSE (0.379) and highest R^2 (0.930) under Scenario 6, though again the difference was only slight compared to the recurrent models.

Table 7.2: Overall Performance on Multi Phase Dataset

Architecture	Metric	Original	1	2	3	4	5	6
LSTM	RMSE	0.376	0.618	0.407	0.403	0.509	0.512	0.383
	MAE	0.203	0.370	0.241	0.234	0.243	0.308	0.217
	R^2	0.930	0.802	0.896	0.917	0.870	0.864	0.929
GRU	RMSE	0.376	0.617	0.405	0.404	0.514	0.512	0.380
	MAE	0.210	0.361	0.238	0.243	0.260	0.315	0.208
	R^2	0.930	0.802	0.897	0.917	0.868	0.864	0.930
Stacked LSTM	RMSE	0.378	0.612	0.407	0.405	0.514	0.514	0.383
	MAE	0.203	0.361	0.240	0.231	0.251	0.320	0.221
	R^2	0.929	0.805	0.896	0.916	0.868	0.863	0.929
Stacked GRU	RMSE	0.380	0.613	0.407	0.407	0.511	0.514	0.386
	MAE	0.207	0.356	0.234	0.243	0.256	0.313	0.234
	R^2	0.928	0.805	0.895	0.916	0.869	0.863	0.928
LSTM + Attention	RMSE	0.378	0.614	0.406	0.406	0.510	0.512	0.382
	MAE	0.221	0.360	0.237	0.245	0.240	0.311	0.198
	R^2	0.929	0.804	0.896	0.916	0.870	0.864	0.929
GRU + Attention	RMSE	0.376	0.614	0.406	0.405	0.510	0.514	0.384
	MAE	0.212	0.356	0.239	0.239	0.254	0.315	0.216
	R^2	0.930	0.804	0.896	0.917	0.870	0.863	0.928
Transformer	RMSE	0.374	0.615	0.405	0.400	0.509	0.510	0.379
	MAE	0.199	0.374	0.218	0.227	0.252	0.317	0.212
	R^2	0.931	0.804	0.897	0.918	0.870	0.866	0.930

7.1.3 Results on the Controlled Cohort Dataset

The controlled cohort dataset (Table 7.3) revealed similar trends but with slightly more pronounced differences between scenarios. Scenario 1 again showed the greatest decline in performance, with R^2 values dropping to the 0.74-0.75 range. Scenarios 2 and 3 were stable, with R^2 exceeding 0.90, confirming that models can handle skewed or temperature-scaled perturbations.

Step-size banding (Scenario 4) and variance inflation (Scenario 5) produced more substantial increases in RMSE, particularly for Stacked models, reflecting difficulty in adapting to abrupt or noisy dynamics. Once again, Scenario 6 maintained high accuracy, with the Transformer achieving $R^2 = 0.932$, nearly matching baseline performance.

Table 7.3: Overall Performance on Controlled Cohort Dataset

Architecture	Metric	Original	1	2	3	4	5	6
LSTM	RMSE	0.350	0.676	0.370	0.397	0.532	0.558	0.376
	MAE	0.187	0.411	0.199	0.217	0.272	0.342	0.198
	R^2	0.937	0.751	0.904	0.917	0.859	0.829	0.929
GRU	RMSE	0.348	0.678	0.371	0.395	0.530	0.556	0.376
	MAE	0.181	0.403	0.213	0.214	0.251	0.346	0.191
	R^2	0.937	0.749	0.903	0.917	0.860	0.830	0.928
Stacked LSTM	RMSE	0.354	0.679	0.372	0.399	0.535	0.561	0.380
	MAE	0.202	0.412	0.199	0.224	0.275	0.352	0.214
	R^2	0.935	0.748	0.903	0.916	0.858	0.827	0.927
Stacked GRU	RMSE	0.352	0.683	0.373	0.399	0.531	0.562	0.378
	MAE	0.198	0.415	0.213	0.220	0.259	0.367	0.198
	R^2	0.936	0.746	0.902	0.916	0.860	0.826	0.928
LSTM + Attention	RMSE	0.350	0.681	0.373	0.398	0.530	0.557	0.379
	MAE	0.190	0.406	0.205	0.216	0.260	0.347	0.207
	R^2	0.937	0.747	0.903	0.916	0.860	0.829	0.927
GRU + Attention	RMSE	0.353	0.676	0.370	0.398	0.531	0.559	0.376
	MAE	0.175	0.414	0.206	0.222	0.258	0.346	0.198
	R^2	0.936	0.750	0.904	0.916	0.860	0.828	0.929
Transformer	RMSE	0.346	0.674	0.367	0.396	0.525	0.553	0.367
	MAE	0.180	0.397	0.204	0.208	0.252	0.334	0.192
	R^2	0.938	0.752	0.906	0.917	0.864	0.832	0.932

7.1.4 Comparison Across Datasets

Overall, performance was consistently strongest in the scenario 6 with structural graph modifications, with R^2 values exceeding 0.91 across all datasets and architectures. Scenario 1, which introduced the strongest distributional shift, consistently produced the most severe degradation in predictive accuracy across all datasets and architectures, underscoring the reliance of temporal models on informative conditional dependencies. Scenarios 2 and 3 were comparatively well handled, demonstrating the models robustness to skewed or smoothed probability distributions. Scenarios 4 and 5 revealed sensitivity to abrupt changes and inflated variance, which introduced greater uncertainty into predictions. Across architectures, differences were modest: LSTM and GRU families performed similarly, with stacked variants showing slightly worse generalisation.

7.2 Item-Level Analysis

While overall model performance provides a useful global assessment, ALS progression is inherently multidimensional, with distinct patterns of decline across functional domains. To gain a finer understanding of this variability, we analysed prediction performance at the level of individual ALSFRS items (Q1-Q10). Item-specific analysis reveals how perturbations affect different domains. Results are reported for the Transformer model, which achieved the strongest performance in most scenarios across different datasets. Metrics include RMSE and R^2, and we report results across all simulation scenarios.

7.2.1 Results on the Single Phase Dataset

Tables 7.4 and 7.5 present Transformer model functional item performance on the single-phase dataset. When perturbations were applied to all functional items (Table 7.4), Scenario 1 consistently degraded performance across Q1-Q10, with RMSE values nearly doubling in some cases (e.g., Q1: $0.380 \rightarrow 0.734$; Q3: $0.404 \rightarrow 0.783$) and R^2 falling below 0.70 for several items. Scenarios 4 and 5 also led to elevated error and reduced explained variance, reflecting the destabilising effect of abrupt shifts and variance amplification.

In contrast, Scenarios 2 and 3 were more benign, with most functional items maintaining R^2 values above 0.83. Scenario 6 proved relatively stable, producing performance close to the original baseline, particularly for Q1, Q8, and Q9.

When perturbations were restricted to Q1, Q2, Q3, and Q10 (Table 7.5), degradation in model performance was more localised. Items outside the perturbed set (e.g., Q4-Q9) retained near-baseline accuracy, with several showing R^2 exceeding 0.85 across scenarios. This confirms that non-stationarity effects are strongest in directly targeted items, while spillover to others remains limited.

Table 7.4: Transformer model: Item Performance on Single Phase Dataset (Q1-Q10)

Item	Metric	Original	1	2	3	4	5	6
Q1	RMSE	0.380	0.734	0.434	0.418	0.511	0.521	0.400
	R^2	0.930	0.714	0.904	0.909	0.844	0.841	0.915
Q2	RMSE	0.429	0.695	0.529	0.488	0.473	0.552	0.411
	R^2	0.882	0.722	0.848	0.854	0.832	0.793	0.888
Q3	RMSE	0.404	0.783	0.543	0.450	0.456	0.528	0.434
	R^2	0.893	0.677	0.831	0.863	0.826	0.813	0.876
Q4	RMSE	0.461	0.717	0.518	0.480	0.523	0.556	0.450
	R^2	0.891	0.738	0.863	0.874	0.856	0.848	0.899
Q5	RMSE	0.492	0.768	0.524	0.541	0.525	0.592	0.495
	R^2	0.873	0.699	0.858	0.829	0.853	0.812	0.878
Q6	RMSE	0.400	0.657	0.433	0.396	0.522	0.545	0.433
	R^2	0.869	0.699	0.859	0.878	0.782	0.788	0.863
Q7	RMSE	0.442	0.685	0.530	0.483	0.581	0.593	0.486
	R^2	0.879	0.706	0.838	0.855	0.800	0.785	0.859
Q8	RMSE	0.391	0.644	0.458	0.417	0.484	0.545	0.409
	R^2	0.877	0.671	0.839	0.847	0.800	0.767	0.848
Q9	RMSE	0.382	0.715	0.433	0.373	0.558	0.521	0.449
	R^2	0.830	0.594	0.819	0.822	0.741	0.740	0.797
Q10	RMSE	0.384	0.748	0.521	0.453	0.454	0.492	0.406
	R^2	0.888	0.689	0.810	0.835	0.830	0.808	0.871

Table 7.5: Transformer model: Item Performance on Single Phase Dataset (Q1, Q2, Q3, Q10 perturbed)

Item	Metric	Original	1	2	3	4	5	6
Q1	RMSE	0.380	0.716	0.456	0.425	0.503	0.511	0.441
	R^2	0.930	0.744	0.900	0.905	0.867	0.853	0.893
Q2	RMSE	0.429	0.799	0.511	0.506	0.514	0.585	0.453
	R^2	0.882	0.682	0.860	0.864	0.832	0.793	0.891
Q3	RMSE	0.404	0.754	0.481	0.467	0.468	0.494	0.449
	R^2	0.893	0.685	0.855	0.850	0.832	0.824	0.857
Q4	RMSE	0.461	0.481	0.431	0.484	0.443	0.476	0.456
	R^2	0.891	0.874	0.902	0.877	0.905	0.885	0.898
Q5	RMSE	0.492	0.519	0.482	0.482	0.520	0.496	0.492
	R^2	0.873	0.863	0.882	0.877	0.858	0.877	0.882
Q6	RMSE	0.400	0.398	0.403	0.404	0.400	0.421	0.408
	R^2	0.869	0.885	0.881	0.878	0.881	0.872	0.882
Q7	RMSE	0.442	0.490	0.460	0.473	0.442	0.478	0.465
	R^2	0.879	0.864	0.878	0.861	0.879	0.862	0.875
Q8	RMSE	0.391	0.430	0.414	0.411	0.410	0.427	0.402
	R^2	0.877	0.836	0.854	0.853	0.857	0.844	0.861
Q9	RMSE	0.382	0.414	0.431	0.374	0.355	0.404	0.369
	R^2	0.830	0.804	0.790	0.855	0.863	0.853	0.847
Q10	RMSE	0.384	0.719	0.556	0.450	0.501	0.503	0.446
	R^2	0.888	0.692	0.803	0.863	0.798	0.806	0.849

7.2.2 Results on the Multi Phase Dataset

Tables 7.6 and 7.7 display results for the multi-phase dataset. Here, Scenarios 2 and 3 were particularly well handled, with R^2 exceeding 0.89 across most functional items. Scenario 1 again showed the strongest degradation, though less severe than in the single phase dataset, suggesting that a larger dataset with additional temporal context is beneficial for the model training.

Interestingly, Scenario 6 (structural modifications) achieved the highest stability, with several items (Q1, Q2, Q9) even surpassing the baseline in R^2. When perturbations were limited to Q1, Q2, Q3, and Q10, non-targeted items remained largely unaffected, confirming that models are not affected by the locality of perturbation effects.

Table 7.6: Transformer model: Item Performance on Multi Phase Dataset (Q1-Q10)

Item	Metric	Original	1	2	3	4	5	6
Q1	RMSE	0.335	0.612	0.394	0.361	0.508	0.475	0.309
	R^2	0.949	0.826	0.920	0.941	0.873	0.893	0.957
Q2	RMSE	0.396	0.641	0.452	0.416	0.513	0.514	0.376
	R^2	0.923	0.797	0.892	0.918	0.844	0.864	0.929
Q3	RMSE	0.390	0.618	0.459	0.406	0.494	0.487	0.402
	R^2	0.924	0.814	0.896	0.919	0.858	0.875	0.908
Q4	RMSE	0.403	0.622	0.412	0.420	0.518	0.520	0.421
	R^2	0.914	0.788	0.875	0.901	0.867	0.850	0.908
Q5	RMSE	0.471	0.680	0.447	0.548	0.549	0.591	0.474
	R^2	0.879	0.746	0.840	0.834	0.851	0.809	0.862
Q6	RMSE	0.320	0.566	0.324	0.326	0.518	0.488	0.341
	R^2	0.913	0.743	0.873	0.905	0.842	0.819	0.912
Q7	RMSE	0.390	0.599	0.402	0.407	0.499	0.519	0.394
	R^2	0.910	0.784	0.867	0.899	0.859	0.841	0.916
Q8	RMSE	0.348	0.610	0.347	0.352	0.478	0.525	0.388
	R^2	0.909	0.729	0.894	0.903	0.831	0.797	0.874
Q9	RMSE	0.308	0.597	0.298	0.309	0.515	0.503	0.297
	R^2	0.879	0.674	0.820	0.858	0.818	0.750	0.920
Q10	RMSE	0.347	0.593	0.470	0.403	0.494	0.458	0.348
	R^2	0.925	0.801	0.875	0.905	0.863	0.873	0.921

Table 7.7: Transformer model: Item Performance on Multi Phase Dataset (Q1, Q2, Q3, Q10 perturbed)

Item	Metric	Original	1	2	3	4	5	6
Q1	RMSE	0.335	0.594	0.393	0.364	0.512	0.485	0.313
	R^2	0.949	0.835	0.922	0.943	0.876	0.888	0.957
Q2	RMSE	0.396	0.606	0.447	0.420	0.508	0.506	0.367
	R^2	0.923	0.818	0.894	0.918	0.859	0.864	0.930
Q3	RMSE	0.390	0.633	0.549	0.411	0.505	0.496	0.402
	R^2	0.924	0.802	0.899	0.922	0.856	0.867	0.912
Q4	RMSE	0.403	0.400	0.403	0.385	0.401	0.400	0.393
	R^2	0.914	0.913	0.915	0.919	0.918	0.916	0.919
Q5	RMSE	0.471	0.512	0.544	0.505	0.459	0.476	0.474
	R^2	0.879	0.857	0.838	0.860	0.887	0.875	0.879
Q6	RMSE	0.320	0.325	0.319	0.320	0.317	0.318	0.318
	R^2	0.913	0.917	0.917	0.917	0.914	0.913	0.915
Q7	RMSE	0.390	0.376	0.393	0.388	0.395	0.386	0.383
	R^2	0.910	0.917	0.906	0.911	0.908	0.913	0.914
Q8	RMSE	0.348	0.345	0.343	0.343	0.329	0.333	0.340
	R^2	0.909	0.911	0.913	0.909	0.915	0.916	0.910
Q9	RMSE	0.308	0.313	0.300	0.295	0.311	0.306	0.297
	R^2	0.879	0.863	0.894	0.886	0.873	0.876	0.873
Q10	RMSE	0.347	0.610	0.465	0.407	0.486	0.469	0.352
	R^2	0.925	0.797	0.877	0.902	0.857	0.855	0.919

7.2.3 Results on the Controlled Cohort Dataset

Finally, Tables 7.8 and 7.9 report results for the controlled cohort dataset. Trends were consistent with the other datasets: Scenario 1 strongly reduced performance, while Scenarios 2 and 3 preserved high accuracy. Scenario 6 consistently provided stable and often near-baseline performance. When perturbations were restricted to the subset Q1, Q2, Q3, and Q10, unaffected items remained close to baseline, confirming the model's robustness in terms of local perturbation effects.

Table 7.8: Transformer model: Item Performance on Controlled Dataset (Q1-Q10)

Item	Metric	Original	1	2	3	4	5	6
Q1	RMSE	0.315	0.649	0.362	0.374	0.526	0.527	0.314
	R^2	0.955	0.784	0.926	0.936	0.859	0.860	0.956
Q2	RMSE	0.361	0.688	0.409	0.410	0.517	0.558	0.377
	R^2	0.935	0.761	0.904	0.921	0.859	0.829	0.932
Q3	RMSE	0.347	0.668	0.417	0.407	0.502	0.521	0.366
	R^2	0.941	0.777	0.910	0.924	0.874	0.851	0.928
Q4	RMSE	0.368	0.695	0.367	0.406	0.512	0.564	0.396
	R^2	0.922	0.722	0.891	0.905	0.866	0.815	0.909
Q5	RMSE	0.477	0.722	0.385	0.571	0.604	0.615	0.483
	R^2	0.870	0.701	0.864	0.807	0.814	0.777	0.843
Q6	RMSE	0.297	0.620	0.287	0.301	0.538	0.539	0.321
	R^2	0.915	0.704	0.878	0.905	0.815	0.760	0.919
Q7	RMSE	0.359	0.672	0.364	0.374	0.505	0.542	0.379
	R^2	0.914	0.724	0.877	0.903	0.865	0.816	0.917
Q8	RMSE	0.309	0.670	0.343	0.344	0.487	0.597	0.368
	R^2	0.920	0.672	0.895	0.906	0.834	0.728	0.888
Q9	RMSE	0.264	0.671	0.257	0.274	0.547	0.553	0.275
	R^2	0.886	0.590	0.834	0.870	0.780	0.670	0.928
Q10	RMSE	0.320	0.678	0.441	0.420	0.498	0.502	0.347
	R^2	0.936	0.767	0.885	0.903	0.867	0.846	0.924

7.2.4 Summary of Item-Level Insights

In summary, item-level analysis confirms that flattening CPTs (Scenario 1) consistently imposes the heaviest performance penalty across datasets, with

Table 7.9: Transformer model: Item Performance on Controlled Dataset (Q1, Q2, Q3, Q10 perturbed)

Item	Metric	Original	1	2	3	4	5	6
Q1	RMSE	0.315	0.654	0.354	0.364	0.520	0.527	0.303
	R^2	0.955	0.791	0.932	0.941	0.865	0.856	0.958
Q2	RMSE	0.361	0.680	0.417	0.414	0.502	0.575	0.392
	R^2	0.935	0.768	0.902	0.916	0.859	0.818	0.922
Q3	RMSE	0.347	0.669	0.431	0.394	0.508	0.535	0.381
	R^2	0.941	0.779	0.908	0.928	0.867	0.843	0.924
Q4	RMSE	0.368	0.389	0.402	0.389	0.384	0.402	0.379
	R^2	0.922	0.916	0.910	0.914	0.923	0.911	0.919
Q5	RMSE	0.477	0.479	0.516	0.479	0.448	0.481	0.472
	R^2	0.870	0.857	0.839	0.858	0.881	0.861	0.874
Q6	RMSE	0.297	0.303	0.288	0.299	0.295	0.304	0.294
	R^2	0.915	0.919	0.920	0.918	0.919	0.911	0.923
Q7	RMSE	0.359	0.372	0.353	0.372	0.362	0.349	0.361
	R^2	0.914	0.910	0.918	0.910	0.917	0.921	0.917
Q8	RMSE	0.309	0.322	0.315	0.315	0.312	0.310	0.308
	R^2	0.920	0.916	0.919	0.921	0.920	0.924	0.925
Q9	RMSE	0.264	0.279	0.268	0.274	0.274	0.269	0.270
	R^2	0.886	0.863	0.885	0.887	0.873	0.889	0.887
Q10	RMSE	0.320	0.653	0.445	0.419	0.490	0.505	0.358
	R^2	0.936	0.790	0.883	0.909	0.867	0.851	0.923

pronounced deterioration in RMSE and R^2. Skewing and temperature scaling (Scenarios 2 and 3) are well tolerated, while step-size banding and variance inflation (Scenarios 4 and 5) introduce moderate instability. Structural modifications (Scenario 6) exhibit the highest robustness, with performance often approaching or even surpassing baseline. Perturbations applied to specific functional items mainly affect those items, with limited spillover effects, highlighting the model's ability to localise non-stationarity in these experiments.

7.3 Phase-Specific Performance

Beyond overall and item-level evaluation, it is crucial to assess how models perform across different stages of disease progression. To this end, prediction performance was stratified by disease phase (Phase 0: early, Phase 1: intermediate, Phase 2: advanced). This stratification reveals whether predictive accuracy varies systematically across phases and whether certain stages present greater challenges for modelling.

7.3.1 Results on the Multi Phase Dataset

Table 7.10 reports phase-specific performance of the Transformer model on the multi phase dataset. Several trends emerge. First, in scenario 1, which introduces strong non-stationarity, Phase 2 exhibited the weakest performance, with RMSE rising to 0.919 (relative RMSE = 0.681) and R^2 dropping to 0.536. By contrast, Phase 0 predictions were far more accurate (R^2 = 0.909). Scenarios 2-4 produced more balanced results across phases, with Phase 1 generally achieving the highest R^2 values (up to 0.898 in scenario 3). Scenario 6 showed the most stable behaviour across phases, with RMSE values near 0.40 and R^2 consistently above 0.91 for all phases. These results highlight that while early-stage disease is generally easier to model, the Transformer was also capable of robust performance in intermediate and late phases under less severe conditions.

7.3.2 Results on the Controlled Cohort Dataset

Table 7.11 shows the corresponding results for the controlled cohort dataset. Similar to the multi phase dataset, scenario 1 revealed sharp degradation in Phase 2 performance (R^2 = 0.516), while Phase 0 maintained higher accuracy

Table 7.10: Transformer: Phase Performance on Multi Phase Dataset

Scenario	Metric	Phase 0	Phase 1	Phase 2
1	RMSE (rel-RMSE)	0.424 (0.301)	0.710 (0.514)	0.919 (0.681)
	R^2	0.909	0.735	0.536
2	RMSE (rel-RMSE)	0.464 (0.342)	0.475 (0.375)	0.524 (0.480)
	R^2	0.882	0.858	0.768
3	RMSE (rel-RMSE)	0.409 (0.291)	0.443 (0.318)	0.489 (0.356)
	R^2	0.915	0.898	0.872
4	RMSE (rel-RMSE)	0.461 (0.337)	0.528 (0.391)	0.579 (0.430)
	R^2	0.886	0.847	0.814
5	RMSE (rel-RMSE)	0.422 (0.304)	0.545 (0.401)	0.704 (0.523)
	R^2	0.907	0.838	0.725
6	RMSE (rel-RMSE)	0.402 (0.287)	0.402 (0.285)	0.410 (0.290)
	R^2	0.917	0.918	0.915

(R^2 = 0.897). Scenario 2, however, produced an unusual pattern: RMSE values decreased markedly from Phase 0 to Phase 2 (0.492 to 0.208), with Phase 2 yielding the strongest performance (R^2 = 0.856). This suggests that in some controlled settings, later stages of disease can be predicted more reliably, likely due to more monotonic trajectories. Scenario 6 again demonstrated robust performance across all phases, with RMSE values below 0.45 and R^2 above 0.90 consistently. Overall, the controlled cohort results reinforce the finding that non-stationarity disproportionately impacts late-stage predictions, but stability can be achieved depending on how severe the perturbation is.

7.3.3 Summary of Phase-Specific Insights

The phase-specific results demonstrate three key insights. First, advanced disease stages (Phase 2) consistently present greater prediction challenges, particularly under non-stationarity scenarios 1 and 5. Second, intermediate phases often yield the highest R^2, likely reflecting smoother trajectories that balance progression signal with reduced variability. Third, under conditions like in scenario 6, performance converges across all phases, highlighting the models robustness when non-stationary effects are introduced by DBN structural changes.

Table 7.11: Transformer: Phase Performance on Controlled Dataset

Scenario	Metric	Phase 0	Phase 1	Phase 2
1	RMSE (rel-RMSE)	0.439 (0.319)	0.663 (0.513)	0.832 (0.695)
	R^2	0.897	0.736	0.516
2	RMSE (rel-RMSE)	0.492 (0.132)	0.367 (0.131)	0.208 (0.143)
	R^2	0.867	0.868	0.856
3	RMSE (rel-RMSE)	0.433 (0.312)	0.396 (0.300)	0.360 (0.336)
	R^2	0.902	0.909	0.886
4	RMSE (rel-RMSE)	0.464 (0.341)	0.533 (0.383)	0.564 (0.427)
	R^2	0.883	0.852	0.816
5	RMSE (rel-RMSE)	0.438 (0.319)	0.521 (0.392)	0.662 (0.532)
	R^2	0.898	0.845	0.716
6	RMSE (rel-RMSE)	0.430 (0.310)	0.374 (0.275)	0.293 (0.236)
	R^2	0.903	0.924	0.943

7.4 Transition Performance Analysis

To gain deeper insight into model behaviour under dynamic conditions, we evaluated performance at four relative time points around phase transitions ($t-2$, $t-1$, t, $t+1$). This analysis focused on identifying how well the models could anticipate, adapt to, and recover from changes in disease dynamics, particularly in the presence of non-stationarity. Phase transitions were defined as points where patients moved between discrete stages of disease progression, with performance metrics computed relative to these anchors.

At each transition point, predictions were compared against ground truth items, and performance was quantified using RMSE. This approach enabled fine-grained evaluation of temporal sensitivity, as the models ability to capture pre-transition trends ($t-2, t-1$), the transition itself (t), and immediate recovery ($t+1$) could be explicitly assessed. These are the points where the data distribution should be most affected by non-stationarity. While the original scenario provides a useful baseline, the six non-stationary scenarios (each introducing different probability perturbations to the CPTs or model structure) represent more challenging conditions. The analysis therefore emphasises Scenarios 1-6.

In this section, we focus exclusively on the Transformer model, as all proposed architectures exhibited broadly similar results with only minor variability in performance. Moreover, the analysis is conducted on the multi-phase dataset, which represents the most challenging setting due to the heterogeneous disease trajectories of patients.

In Scenario 1 (Figure 7.1), a noticeable increase in RMSE occurs at the transition points, particularly between visits $t-1$ and t. However, the model demonstrates a recovery trend within Phase 1, where RMSE values decline across subsequent visits. This reduction is especially pronounced in question Q1, and a similar pattern is observed for the Phase 1 to Phase 2 transition, where the average RMSE decreases from 0.864 in the first visit to 0.827 in the second.

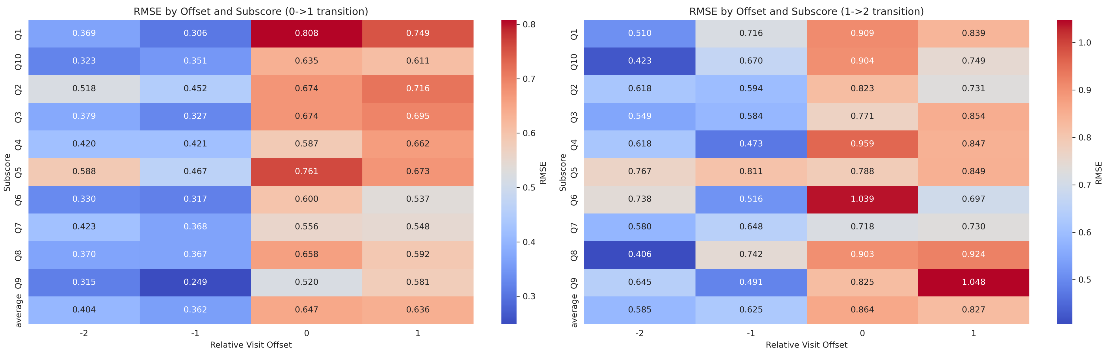

Figure 7.1: Transition Performance of the Transformer model for Scenario 1 on the Multi-phase dataset.

Scenario 2 (Figure 7.2) shows that the model can adapt relatively quickly to extreme skew perturbations. For example, Q3 in Phase 1 exhibits a reduction in RMSE from 0.498 to 0.419 between the first and second visits. Similarly, items Q2 and Q3 improve in Phase 2, with RMSE dropping from 0.574 to 0.389 and from 0.548 to 0.429, respectively. These improvements highlight the robustness of the model to sudden distortions in probability mass.

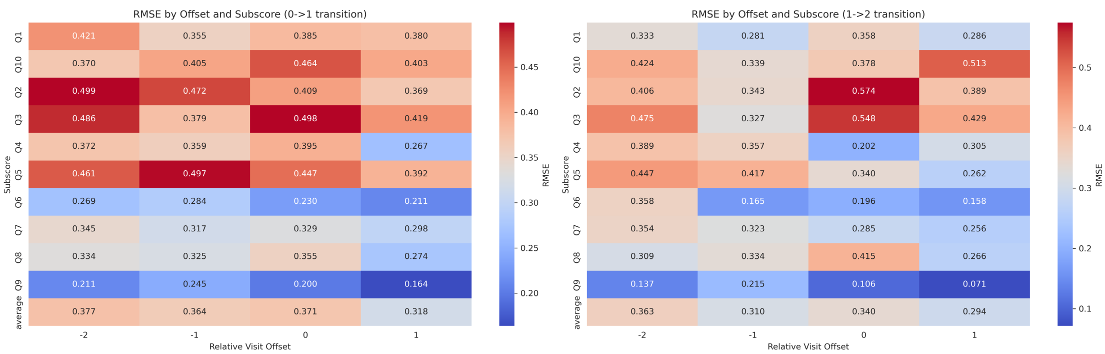

Figure 7.2: Transition Performance of the Transformer model for Scenario 2 on the Multi-phase dataset.

In Scenario 3 (Figure 7.3), the overall RMSE values remain relatively stable across transitions, with no drastic shifts between neighboring visits. Nevertheless, visits within Phase 1 tend to show slightly higher RMSE on average compared to Phases 0 and 2. This suggests that temperature scaling creates a more subtle form of non-stationarity that slightly increases prediction uncertainty without producing sharp discontinuities.

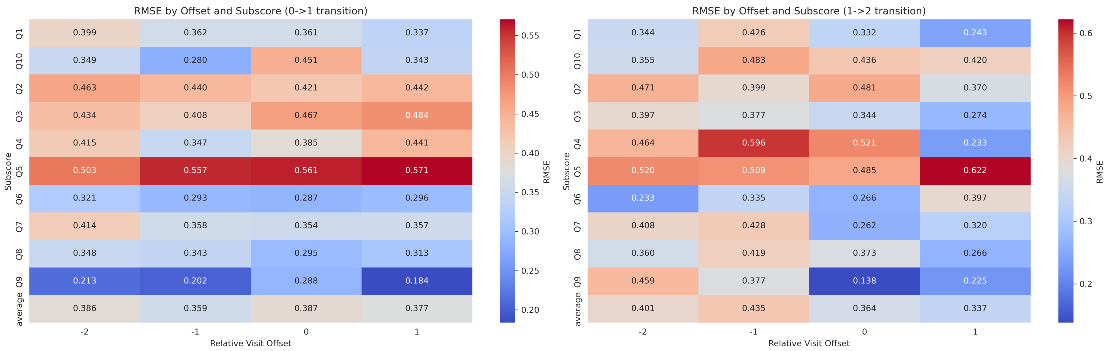

Figure 7.3: Transition Performance of the Transformer model for Scenario 3 on the Multi-phase dataset.

Scenario 4 (Figure 7.4) introduces banded perturbations, leading to sharp but transient increases in RMSE at phase entry points. For example, Q7 and Q9 in the first visit of Phase 1 spike to 0.716 and 0.663, respectively, before dropping to 0.547 and 0.478 in the second visit. Average RMSE follows a similar pattern, decreasing from 0.562 and 0.538 at the start of Phase 1 to 0.478 and 0.498 toward the end. A comparable effect is observed at the onset of Phase 2, where items Q4 and Q5 increase sharply to 0.673 and 0.820 but decline to 0.625 and 0.545 by the next visit. These dynamics indicate that the model can recover after abrupt perturbations but requires one or two visits to stabilise.

Variance inflation (Figure 7.5) proves more challenging for the Transformer model. Here, increases in RMSE at transition points are not followed by consistent recovery, and overall performance remains degraded across visits. This scenario highlights a limitation of the model architecture when confronted with distributions that exhibit inflated variability, as the uncertainty propagates more strongly into the predictions.

Scenario 6 (Figure 7.6) involves structural changes to the CPTs, producing sharp increases in RMSE at the first visits of new phases. Questions such as Q2, Q3, Q6, and Q7 show initial spikes (0.678, 0.564, 0.579, and 0.663, respectively), followed by immediate drops in the next visit (0.424, 0.444, 0.331, and 0.409).

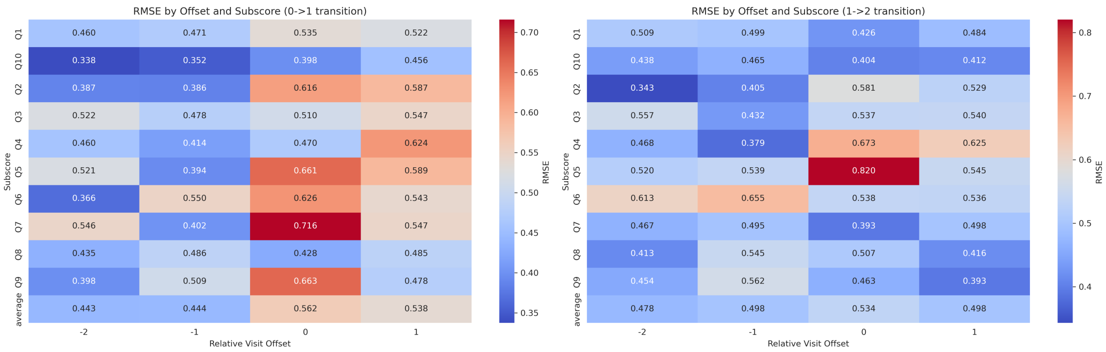

Figure 7.4: Transition Performance of the Transformer model for Scenario 4 on the Multi-phase dataset.

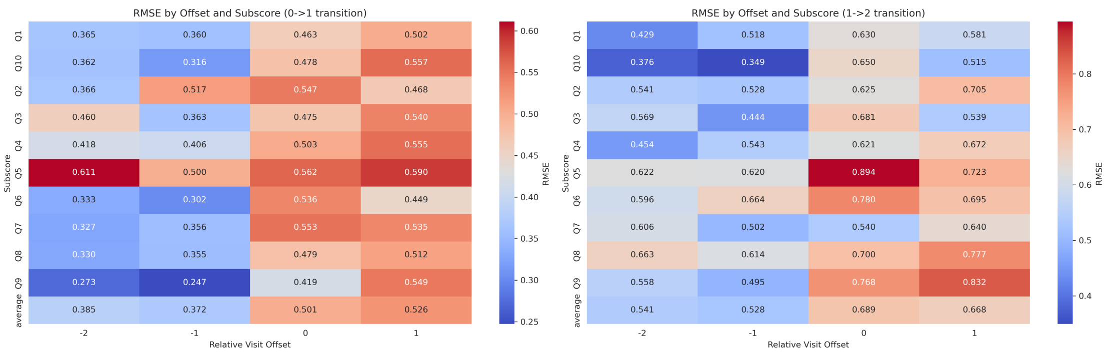

Figure 7.5: Transition Performance of the Transformer model for Scenario 5 on the Multi-phase dataset.

The same tendency is visible at the Phase 1 to Phase 2 transition, where Q5 and Q6 rise to 0.633 and 0.431 before decreasing to 0.464 and 0.203, respectively. Average RMSE across all items mirrors this pattern, decreasing from 0.387 in the first visit of Phase 2 to 0.331 in the second. These results suggest that the model is relatively resilient to structural changes, rapidly restabilising after abrupt shocks.

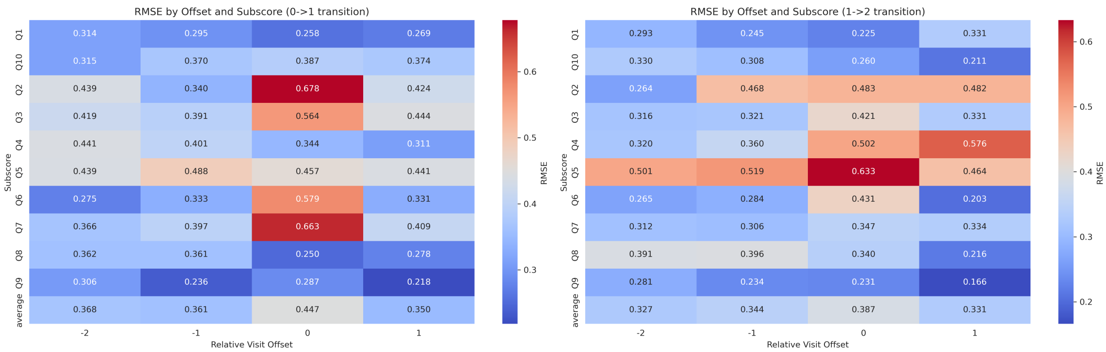

Figure 7.6: Transition Performance of the Transformer model for Scenario 6 on the Multi-phase dataset.

Across all scenarios, the results indicate that while non-stationarity introduces temporary increases in prediction error, the Transformer architecture exhibits notable robustness in most cases. Scenarios involving flattening, skewing, or structural modifications show that the model can start adapting within one or two visits. However, under variance inflation (Scenario 5), recovery is limited, suggesting that scenarios involving heightened uncertainty pose the greatest challenge. Overall, these findings demonstrate that while the model is sensitive to abrupt distributional shifts, its capacity for recovery varies depending on the type of perturbation applied.

7.5 Comparative Analysis Across Models and Scenarios

7.5.1 Key Findings Across Architectures

Overall, the results demonstrate that RNN families (LSTM and GRU) and their stacked or attention-augmented variants exhibit broadly similar behaviour across scenarios. Differences between individual architectures were modest, typically amounting to small fluctuations in RMSE or R^2 that did not alter the general trends. Stacked models tended to display slightly higher error under challenging scenarios, suggesting reduced generalisation when model depth is increased without a corresponding expansion of training data. The Transformer consistently delivered stable performance, often achieving the lowest RMSE and highest R^2 across scenarios, but the advantage relative to recurrent baselines was incremental rather than transformative.

These findings confirm the initial hypothesis that temporal deep learning models, regardless of specific architecture, are capable of learning the structure of disease progression when trained on large-scale synthetic datasets. The similarities across models indicate that robustness to non-stationarity is more strongly influenced by the nature of the perturbation than by the choice of architecture.

7.5.2 Comparison Across Scenarios

The comparative analysis across scenarios highlights the critical role of distributional shifts in shaping model behaviour. Scenario 1 (flattening CPTs) consistently imposed the largest performance penalty across all models and datasets, underscoring their reliance on informative temporal dependencies. By contrast, Scenarios 2 (extreme skewing) and 3 (temperature scaling) were well tolerated, with models rapidly adapting and maintaining high explained variance. This suggests that the architectures are robust to skewed or smoothed distributions, as long as the underlying temporal structure remains intact.

Scenarios 4 (step-size banding) and 5 (variance inflation) revealed greater sensitivity, particularly at transition points between disease phases. While models could partially recover from banding perturbations, variance inflation produced persistent increases in error, indicating that inflated uncertainty propagates through the models and degrades stability. Scenario 6 (structural modifications) emerged as the most benign non-stationary setting, with performance nearly indistinguishable from the original baseline in many cases. The results suggest that while altered dependency structures require readaptation, the models can quickly establish new predictive patterns.

7.5.3 Comparison Across Datasets

Dataset characteristics also shaped performance outcomes. The multi-phase dataset consistently yielded stronger results than the single-phase dataset, with lower RMSE and higher R^2. This finding supports the expectation that larger and more diverse training sets provide richer temporal context, allowing models to better capture disease progression dynamics. The controlled cohort dataset, though smaller, highlighted sharper contrasts across scenarios, revealing that reduced variability in patient trajectories makes perturbations more visible in performance metrics.

These findings align with the research objectives of assessing robustness to non-stationarity under varying conditions. They demonstrate that while larger datasets confer stability, the type of perturbation remains the dominant factor in determining predictive accuracy.

8

Conclusion and Future Work

8.1 Summary of Research Objectives and Findings

The primary objective of this thesis was to investigate deep sequence modelling approaches for non-stationary disease progression, using ALS as a motivating case study. To achieve this, a simulation framework was developed based on DBNs to generate synthetic longitudinal datasets with controlled non-stationarity. Within this framework, several scenarios were designed to perturb disease dynamics, enabling systematic evaluation of predictive models under shifting conditions.

Deep sequential architectures, including LSTM, GRU, and Transformer models, were trained and evaluated across datasets, with a particular emphasis on the multi-phase setting where patients follow heterogeneous disease trajectories. A transition-aware evaluation framework was introduced to assess predictive accuracy around clinically relevant progression points.

The findings demonstrate that all architectures were able to adapt to non-stationarity, though with varying degrees of robustness. The Transformer consistently exhibited strong performance across scenarios, particularly in adapting quickly to probability perturbations. These results highlight the utility of combining simulation-based benchmarking with deep sequence modelling for addressing complex temporal dynamics in disease progression.

8.2 Implications for ALS Research

The contributions of this work extend beyond methodological advances and hold relevance for the broader ALS research community. First, the simulation framework provides a novel way to generate synthetic patient trajectories, which can be used to augment clinical studies where data scarcity or missingness pose significant challenges. Second, the results emphasise the importance of modelling ALSFRS items rather than aggregate scores, as this enables more fine-grained predictions of patient function across disease phases.

More generally, the thesis contributes to the growing body of work on non-stationary sequence modelling. The proposed evaluation framework illustrates how models behave under different types of distribution shifts, a perspective applicable not only to ALS but also to other (chronic) diseases characterised by progressive decline. The use of synthetic data for benchmarking further demonstrates how controlled experiments can guide the development of clinically meaningful predictive models.

8.3 Future Directions

While the thesis provides a solid foundation for studying non-stationary disease progression, several avenues for future research remain.

- **Complex Non-Stationary Scenarios:** Extending the simulation framework to capture richer forms of non-stationarity, such as multi-factor perturbations, or irregular visit schedules, would provide a more comprehensive stress test of predictive models.

- **Alternative Architectures:** Exploring newer architectures, including diffusion models, or hybrid probabilistic-deep learning approaches, could enhance robustness to temporal variability and provide interpretable predictions.

- **Integration with Real-World Data:** Validating the synthetic findings on external ALS cohorts or registries could strengthen the clinical utility of the framework and provide insights into model generalisation.

- **Patient-Specific Personalisation:** Investigating transfer learning or meta-learning techniques to adapt models at the individual patient level may lead to improved personalised progression predictions.

8.4 Final Remarks

This thesis introduced a novel framework for simulating and predicting non-stationary disease progression, with a focus on ALS as a motivating application. By integrating probabilistic simulation with deep sequence modelling, it addressed important methodological and clinical gaps, advancing both the design of benchmarking datasets and the evaluation of temporal prediction models. The work highlights several strengths, including the systematic treatment of non-stationarity, the development of phase-aware evaluation metrics, and the demonstration of robust Transformer performance across scenarios. At the same time, it acknowledges limitations such as reliance on synthetic datasets, simplified non-stationary perturbations, and the need for external validation.

Overall, this thesis contributes to the broader field of computational disease modelling by proposing tools and methodologies that can inform both ALS research and the study of other chronic conditions. It reinforces the importance of developing predictive models capable of adapting to the dynamic and heterogeneous nature of disease, offering a pathway towards more robust and clinically meaningful progression modelling in the future.

References

[1] Kingma DP Ba J Adam et al. "A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* 1412.6 (2014).

[2] Ahmed M Alaa and Mihaela van der Schaar. "Forecasting individualized disease trajectories using interpretable deep learning". In: *arXiv preprint arXiv:1810.10489* (2018).

[3] V Appel et al. "A rating scale for amyotrophic lateral sclerosis: description and preliminary experience". In: *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 22.3 (1987), pp. 328–333.

[4] Nazem Atassi et al. "The PRO-ACT database: design, initial analyses, and predictive features". In: *Neurology* 83.19 (2014), pp. 1719–1725.

[5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. "Layer normalization". In: *arXiv preprint arXiv:1607.06450* (2016).

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (2014).

[7] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv". In: *arXiv preprint arXiv:1803.01271* 10 (2018).

[8] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. "Learning long-term dependencies with gradient descent is difficult". In: *IEEE transactions on neural networks* 5.2 (1994), pp. 157–166.

[9] Benjamin Rix Brooks et al. "El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis". In: *Amyotrophic lateral sclerosis and other motor neuron disorders* 1.5 (2000), pp. 293–299.

[10] Jesse M Cedarbaum and Nancy Stambler. "Performance of the amyotrophic lateral sclerosis functional rating scale (ALSFRS) in multicenter clinical trials". In: *Journal of the neurological sciences* 152 (1997), s1–s9.

[11] Jesse M Cedarbaum et al. "The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function". In: *Journal of the neurological sciences* 169.1-2 (1999), pp. 13–21.

[12] Zhengping Che et al. "Recurrent neural networks for multivariate time series with missing values". In: *Scientific reports* 8.1 (2018), p. 6085.

[13] Qi Chen et al. "Dynamic Bayesian network for predicting physiological changes, organ dysfunctions and mortality risk in critical trauma patients". In: *BMC Medical Informatics and Decision Making* 22.1 (2022), p. 119.

[14] David Maxwell Chickering. "Optimal structure identification with greedy search". In: *Journal of machine learning research* 3.Nov (2002), pp. 507–554.

[15] Adriano Chio et al. "Prognostic factors in ALS: a critical review". In: *Amyotrophic lateral sclerosis* 10.5-6 (2009), pp. 310–323.

[16] Adriano Chiò et al. "Development and evaluation of a clinical staging system for amyotrophic lateral sclerosis". In: *Journal of Neurology, Neurosurgery & Psychiatry* 86.1 (2015), pp. 38–44.

[17] Junyoung Chung et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: *arXiv preprint arXiv:1412.3555* (2014).

[18] Thomas Dean and Keiji Kanazawa. "A model for reasoning about persistence and causation". In: *Computational intelligence* 5.2 (1989), pp. 142–150.

[19] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society: series B (methodological)* 39.1 (1977), pp. 1–22.

[20] Muzammil Arif Din Abdul Jabbar et al. "Predicting amyotrophic lateral sclerosis (ALS) progression with machine learning". In: *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* 25.3-4 (2024), pp. 242–255.

[21] *Dive into Deep Learning (D2L.ai)*. `https://d2l.ai/`. Accessed: 2025-10-01. 2025.

[22] Ton Fang et al. "Comparison of the Kings and MiToS staging systems for ALS". In: *Amyotrophic lateral sclerosis and frontotemporal degeneration* 18.3-4 (2017), pp. 227–232.

[23] Eleonora Fresi. "Predictive models for Amyotrophic Lateral Sclerosis disease progression: assessment and improvement of a state-of-the-art model". Relatore: Dott.ssa Martina Vettoretti. Correlatori: Prof.ssa Barbara Di Camillo, Dott.ssa Erica Tavazzi. Master's Thesis. Università degli Studi di Padova, Dipartimento di Ingegneria dellInformazione, Dec. 2021. URL: `http://tesi.cab.unipd.it/`.

[24] Marzyeh Ghassemi et al. "A review of challenges and opportunities in machine learning for health". In: *AMIA Summits on Translational Science Proceedings* 2020 (2020), p. 191.

[25] Alessandro Guazzo et al. "Predicting clinical events characterizing the progression of amyotrophic lateral sclerosis via machine learning approaches using routine visits data: a feasibility study". In: *BMC Medical Informatics and Decision Making* 24.Suppl 4 (2024), p. 318.

[26] Orla Hardiman et al. "Amyotrophic lateral sclerosis". In: *Nature reviews Disease primers* 3.1 (2017), pp. 1–19.

[27] Hrayr Harutyunyan et al. "Multitask learning and benchmarking with clinical time series data". In: *Scientific data* 6.1 (2019), p. 96.

[28] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[29] Torsten Hothorn and Hans H Jung. "RandomForest4Life: a random forest for predicting ALS disease progression". In: *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* 15.5-6 (2014), pp. 444–452.

[30] Matthew C Kiernan et al. "Amyotrophic lateral sclerosis". In: *The lancet* 377.9769 (2011), pp. 942–955.

[31] Neville Kenneth Kitson et al. "A survey of Bayesian Network structure learning". In: *Artificial Intelligence Review* 56.8 (2023), pp. 8721–8814.

[32] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[33] Vyacheslav Kungurtsev et al. "Learning Dynamic Bayesian Networks from Data: Foundations, First Principles and Numerical Comparisons". In: *arXiv preprint arXiv:2406.17585* (2024).

[34] V La Bella et al. "Metodi di valutazione della progressione di malattia nella sclerosi laterale amiotrofica". In: *Neurol sci* 25 (2004), S61–S64.

[35] Yikuan Li et al. "BEHRT: transformer for electronic health records". In: *Scientific reports* 10.1 (2020), p. 7155.

[36] Giancarlo Logroscino et al. "Incidence of amyotrophic lateral sclerosis in Europe". In: *Journal of Neurology, Neurosurgery & Psychiatry* 81.4 (2010), pp. 385–390.

[37] Paul Mehta. "Prevalence of amyotrophic lateral sclerosisUnited States, 2015". In: *MMWR. Morbidity and mortality weekly report* 67 (2018).

[38] TL Munsat et al. "The natural history of motoneuron loss in amyotrophic lateral sclerosis". In: *Neurology* 38.3 (1988), pp. 409–409.

[39] Kevin Patrick Murphy. *Dynamic bayesian networks: representation, inference and learning*. University of California, Berkeley, 2002.

[40] Andrew A Neath and Joseph E Cavanaugh. "The Bayesian information criterion: background, derivation, and applications". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 4.2 (2012), pp. 199–203.

[41] Corrado Pancotti et al. "Deep learning methods to predict amyotrophic lateral sclerosis disease progression". In: *Scientific reports* 12.1 (2022), p. 13738.

[42] Fabiano Papaiz et al. "Predicting ALS progression using Autoregressive deep learning models". In: *Intelligence-Based Medicine* (2025), p. 100247.

[43] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[44] Haoshen Qin et al. "Optimizing deep learning models to combat amyotrophic lateral sclerosis (ALS) disease progression". In: *Digital health* 11 (2025), p. 20552076251349719.

[45] Jose C Roche et al. "A proposed staging system for amyotrophic lateral sclerosis". In: *Brain* 135.3 (2012), pp. 847–852.

[46] Lewis P Rowland. "How amyotrophic lateral sclerosis got its name: the clinical-pathologic genius of Jean-Martin Charcot". In: *Archives of neurology* 58.3 (2001), pp. 512–515.

[47] Donald B Rubin. “Inference and missing data”. In: *Biometrika* 63.3 (1976), pp. 581–592.

[48] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. “Practical bayesian optimization of machine learning algorithms”. In: *Advances in neural information processing systems* 25 (2012).

[49] Nitish Srivastava et al. “Dropout: a simple way to prevent neural networks from overfitting”. In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[50] Erica Tavazzi et al. “Artificial intelligence and statistical methods for stratification and prediction of progression in amyotrophic lateral sclerosis: A systematic review”. In: *Artificial Intelligence in Medicine* 142 (2023), p. 102588.

[51] Erica Tavazzi et al. “Predicting functional impairment trajectories in amyotrophic lateral sclerosis: a probabilistic, multifactorial model of disease progression”. In: *Journal of neurology* 269.7 (2022), pp. 3858–3878.

[52] The ALS Association. *What is ALS?* `https://www.als.org/understanding-als/what-is-als`. Accessed: 2025-09-30. 2025.

[53] BJ Traynor et al. “Functional outcome measures as clinical trial endpoints in ALS”. In: *Neurology* 63.10 (2004), pp. 1933–1935.

[54] Stef Van Buuren and Karin Groothuis-Oudshoorn. “mice: Multivariate imputation by chained equations in R”. In: *Journal of statistical software* 45 (2011), pp. 1–67.

[55] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[56] Liming Wang et al. “Automatic prediction of amyotrophic lateral sclerosis progression using longitudinal speech transformer”. In: *arXiv preprint arXiv:2406.18625* (2024).

[57] Cao Xiao, Edward Choi, and Jimeng Sun. “Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review”. In: *Journal of the American Medical Informatics Association* 25.10 (2018), pp. 1419–1428.

[58] Jiabao Xu et al. “A survey of deep learning for electronic health records”. In: *Applied Sciences* 12.22 (2022), p. 11709.

[59] Neta Zach et al. "Being PRO-ACTive: what can a clinical trial database reveal about ALS?" In: *Neurotherapeutics* 12.2 (2015), pp. 417–423.

[60] Alessandro Zandona et al. "A dynamic Bayesian network model for the simulation of amyotrophic lateral sclerosis progression". In: *BMC bioinformatics* 20.Suppl 4 (2019), p. 118.

Acknowledgments

I would like to express my heartfelt gratitude to my supervisors, Prof. Barbara Di Camillo and Prof. Erica Tavazzi, for their invaluable guidance in defining the topic of my thesis and for their continuous support throughout the course of my research. Their expertise and constructive feedback were essential in shaping this work.

A special thanks goes to my family and friends back home, whose constant support, even from a distance, has given me the strength to overcome every challenge during these two years. Your encouragement and belief in me have been a source of motivation throughout this journey.

Finally, I would like to thank all the friends and classmates I met during my stay in Padova for the many memories we created together, for their collaboration and stimulating discussions, and for contributing to an inspiring and supportive academic environment.