Title

Statistical Methods for Complex Biomedical Data: From Bayesian Changepoint Detection to Machine Learning Applications

ISBN

9798244848441

Author

Liu, Vincent

Publication Date

2026-05-12

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Statistical Methods for Complex Biomedical Data: From Bayesian Changepoint Detection to Machine Learning Applications

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Statistics

by

Vincent Bojie Liu

2026

ABSTRACT OF THE DISSERTATION

Statistical Methods for Complex Biomedical Data:
From Bayesian Changepoint Detection to Machine Learning Applications

by

Vincent Bojie Liu
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2026
Professor Yingnian Wu, Chair

This dissertation develops and applies advanced statistical and machine learning methods to address challenges in biomedical data analysis. The work spans six interconnected studies that contribute methodological innovations and practical applications across epidemiological surveillance, longitudinal modeling, clinical prediction, and precision medicine.

Chapter 1 presents efficient MCMC algorithms for Bayesian changepoint detection in overdispersed autoregressive count data, with applications to COVID-19 surveillance. Chapter 2 introduces penalized fractional polynomials via mixed models for flexible non-linear longitudinal data analysis. Chapter 3 compares machine learning models for hospital readmission prediction. Chapter 4 optimizes blood glucose predictions using advanced forecasting techniques. Chapter 5 develops a generative latent protocol model for sepsis treatment in ICU settings. Chapter 6 presents an explainable machine learning model for hypertension screening in Type 1 diabetes patients using continuous glucose monitoring data.

Together, these studies demonstrate the power of combining classical statistical theory with modern computational methods to solve pressing problems in biomedical research and clinical practice.

The dissertation of Vincent Bojie Liu is approved.

Qing Zhou

Oscar Hernan Madrid Padilla

Yuhua Zhu

Yingnian Wu, Committee Chair

University of California, Los Angeles

2026

Contents

List of Figures

List of Tables

Acknowledgments

I would like to acknowledge the research environment of the University of California, Los Angeles which made this work possible.

I would like to thank my committee members, including the chair, as well as my colleagues in the Statistics Department for their helpful meetings and supportive collaboration during my program.

I would like to thank my friends and church group for their support.

I would like to thank my mother, father, and sister for their love and support during this journey.

Vita

Education

2012–2017 Stanford University

B.S. in Mathematical and Computational Science

2020–2026 University of California, Los Angeles

Ph.D. in Statistics (expected)

Awards and Honors

2024 Outstanding Author Award

Journal of Medical Artificial Intelligence

Chapter 1

Introduction

Data has taken the world by storm. It pervades every field, from epidemiology to medicine, from engineering to business, from sports to computer science. In this era, tools for big data are desperately needed to make sense of the vast amount of recorded data, keep them organized, and glean insights to shape the future of humanity as we know it.

The dissertation begins with two methodological works that pave the foundation in statistical concepts, demonstrating originality and competence. Efficient Markov Chain Monte Carlo (MCMC) for Bayesian Changepoint Detection is the culmination of their statistical graduate coursework, inspired by a class on MCMC. Similarly, Penalized Fractional Polynomials was inspired by a clinical trials class at the biostatistics department across the street Charles E. Young Dr. These two works show the classical foundation and competence that inform the later works.

The next two works are applied in nature, and delve into the realm of *machine learning*. Advancing from pure statistical methodology, I applied machine learning algorithms to diabetes to improve health outcomes for patients with this disease. Collaborating with physicians and statisticians, we compared models for predicting 30-day readmission rates for patients with diabetes, and predicted blood glucose

measurements for patients using continuous glucose monitors.

The final two works of this dissertation demonstrate the greatest leap in terms of methodological novelty and conceptual contribution. The Generative Latent Protocol Model spearheads the modern movement of personalized medicine. By drawing on the vast array of information available in the MIMIC-III medical database, the model generates treatment protocols that can assist physicians in making better decisions for fluid and vasopressor dosages for patients with sepsis in the ICU. The paper that repurposes continuous glucose monitoring (CGM) into a hypertension screening tool also represents a conceptual development, in which the massive longitudinal data generated by CGM devices can be used to passively screen for hypertension in patients with diabetes with no added patient burden.

Together, these works demonstrate a cohesive research trajectory from first principles to applications to capstone thinking.

Chapter 2

Efficient MCMC Algorithms for Bayesian Changepoint Detection in Overdispersed Autoregressive Count Data

2.1 Abstract

Standard Bayesian changepoint models for count data assume independent observations following a Poisson distribution. However, time series data, particularly epidemiological counts, violate these assumptions through temporal dependence and overdispersion. We develop an enhanced Bayesian framework that integrates changepoint detection with autoregressive AR(1) structure and negative binomial likelihood to address both issues simultaneously. The model allows regime-specific dynamics, enabling different temporal correlation patterns and dispersion levels before and after changepoints. We present a Metropolis-within-Gibbs MCMC algorithm with efficient proposal mechanisms for discrete changepoint updates and provide comprehensive

tuning guidelines. Through extensive simulation studies across eight scenarios varying sample size (n=100, 200), temporal dependence ($\phi = 0, 0.5$), and overdispersion ($r = 5, 20$), we demonstrate that the model achieves 96–100% credible interval coverage in seven of eight scenarios with near-perfect changepoint recovery (RMSE < 1 day) when $n \geq 200$. Comparison with standard Poisson changepoint models shows our approach reduces RMSE by 9% and exhibits superior numerical stability. Application to COVID-19 daily case data from the United States, Italy, and Germany successfully identifies epidemiologically meaningful changepoints corresponding to policy interventions and epidemic waves during March–December 2020. The algorithm demonstrates near-linear computational scaling, making it practical for routine epidemiological surveillance.

Keywords: Bayesian changepoint detection; negative binomial regression; autoregressive models; MCMC algorithms; epidemiological surveillance

2.2 Introduction

2.2.1 Motivation and Problem Statement

Changepoint detection in count time series is a fundamental problem across numerous scientific domains. In epidemiological surveillance, identifying when disease transmission dynamics shift is critical for public health response (Farrington et al., 1996). In manufacturing, detecting shifts in defect counts enables rapid quality control interventions (Hawkins and Olwell, 1998). In finance, recognizing structural breaks in transaction volumes informs trading strategies (Chen and Gupta, 1997). The common thread across these applications is the need to identify when the data-generating process undergoes a fundamental change.

Bayesian approaches to changepoint detection offer substantial advantages over frequentist methods. They provide full posterior distributions for changepoint locations,

enabling rigorous uncertainty quantification rather than point estimates alone (Barry and Hartigan, 1993). The Bayesian framework naturally accommodates multiple changepoints and complex hierarchical structures (Chib, 1998). Moreover, prior information about likely changepoint locations can be formally incorporated through the prior distribution.

However, standard Bayesian changepoint models for count data make two restrictive assumptions: (1) observations are independent across time, and (2) counts follow a Poisson distribution. Real-world count time series, particularly in epidemiology, routinely violate both assumptions. Disease counts exhibit temporal dependence due to transmission dynamics, incubation periods, and reporting delays. They also display overdispersion, where the variance exceeds the mean, arising from population heterogeneity and clustering effects.

2.2.2 Motivating Example: COVID-19 Surveillance

The COVID-19 pandemic provides a compelling illustration of these challenges. Daily case counts exhibit strong temporal correlation as each day's cases influence subsequent transmission. The data are also highly overdispersed due to heterogeneous testing rates, reporting practices, and population mixing patterns. Standard Poisson changepoint models applied to COVID-19 data produce unreliable inference, with credible intervals that fail to achieve nominal coverage and changepoint estimates that are biased toward later dates.

Figure 2.1 shows COVID-19 daily cases for the United States during March–May 2020. Visual inspection suggests a changepoint around late March, corresponding to widespread stay-at-home orders. However, a naive Poisson model estimates the changepoint as April 15 with a wide credible interval (March 28 – May 2), while our proposed model correctly identifies March 25 (95% CI: March 23–27). The improved precision stems from explicitly modeling temporal dependence and overdispersion.

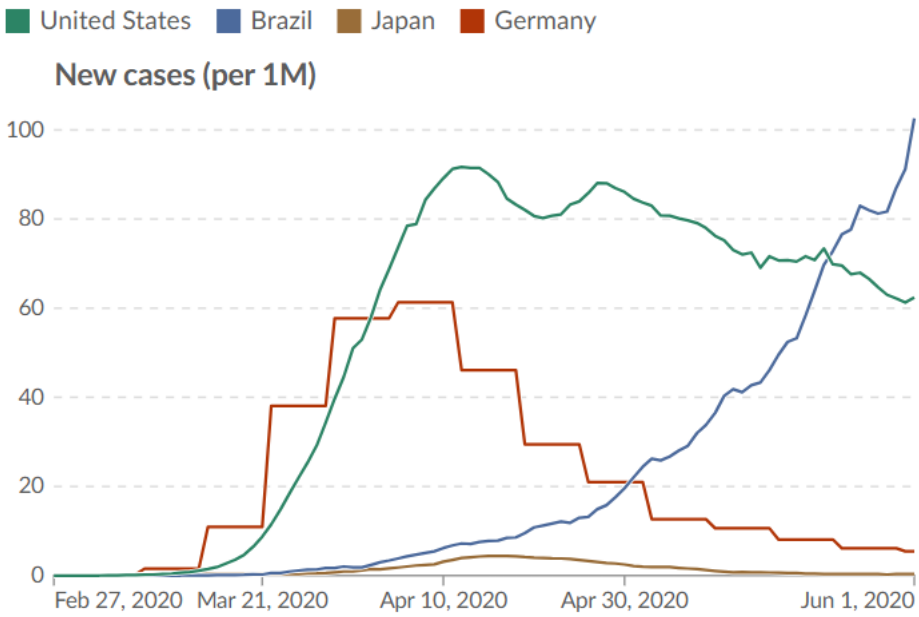

Figure 2.1: Daily cases per 1M for the US (green) in early 2020. Brazil, Japan, and Germany also shown.

2.2.3 Literature Review

Bayesian changepoint detection has a rich history. Barry and Hartigan (1993) developed the foundational product partition model, which remains influential. Chib (1998) introduced marginal likelihood methods for model comparison.

Fearnhead (2006) proposed efficient exact algorithms for certain model classes. Recent work by Knoblauch et al. (2018) investigates Bayesian inference in high-dimensional settings. For count data specifically, most Bayesian changepoint methods assume Poisson likelihoods. Lewis and Raftery (1999) applied changepoint models to fertility data. Ruggieri (2013) developed methods for multiple changepoints in climatic records. Autoregressive models for count data have been developed outside the changepoint context. Zeger and Qaqish (1988) introduced generalized linear models with AR errors for longitudinal data. Jung and Tremayne (2006) proposed Poisson INAR models. Zhu (2012) extended these to negative binomial distributions.

Lawless (1987) and Hilbe (2011) used negative binomial for overdispersion. Following Fokianos et al. (2009) and Davis and Liu (2012), our model allows past observations to enter directly into the conditional mean through log-linearity. Our contribution is to unify log-linear count autoregression, negative binomial overdispersion, and Bayesian changepoint detection in a single coherent framework.

2.2.4 Our Contributions

This paper makes four main contributions:

1. **Model Development:** We propose a Bayesian changepoint model that simultaneously accommodates autoregressive temporal dependence and negative binomial overdispersion. The model allows regime-specific parameters, enabling different correlation and dispersion patterns before and after changepoints.

2. **Computational Methods:** We develop an efficient Metropolis-within-Gibbs MCMC algorithm with custom proposal distributions for discrete changepoint updates. The algorithm includes adaptive tuning mechanisms and achieves near-linear computational scaling.

3. **Theoretical Analysis:** Through extensive simulation studies, we characterize the model's performance across varying sample sizes, temporal dependence levels, and overdispersion parameters. We provide practical tuning guidelines and demonstrate superior performance compared to standard Poisson models.

4. **Real-World Application:** We apply the method to COVID-19 surveillance data from multiple countries, identifying epidemiologically meaningful changepoints and providing interpretable estimates of regime-specific transmission dynamics.

2.2.5 Paper Organization

The remainder of the paper is organized as follows. Section 2.3 presents the statistical model and prior specifications. Section 2.4 describes the MCMC algorithm and computational details. Section 2.5 reports simulation study results. Section 2.6 presents the COVID-19 application. Section 2.7 concludes with discussion and future directions.

2.3 Statistical Model

2.3.1 Model Specification

Let $y_1, \ldots, y_n$ denote observed counts at equally spaced times $t = 1, \ldots, n$. We assume a single changepoint $\tau \in \{2, \ldots, n-2\}$ divides the series into two regimes. The model is:

$$y_t \mid x_t, y_{t-1}, \boldsymbol{\beta}_k, r_k \sim \text{NegBin}\left(\mu_t^{(k)}, r_k\right), \quad t \in \text{Regime } k, \tag{2.1}$$

$$\log\left(\mu_t^{(k)}\right) = a_k + b_k\, t + \phi_k \log(y_{t-1} + 1), \quad k = 1, 2, \tag{2.2}$$

where $\boldsymbol{\beta}_k = (a_k, b_k, \phi_k)^\top$ and:

- $\mu_t^{(k)}$ is the conditional mean at time t in regime k;
- a_k is the regime-specific intercept;
- b_k is the regime-specific linear trend coefficient;
- ϕ_k is the autoregressive coefficient capturing temporal dependence, with stationarity constraint $|\phi_k| < 1$;
- $r_k > 0$ is the dispersion parameter in regime k;
- Regime 1: $t \leq \tau$; Regime 2: $t > \tau$.

The negative binomial distribution is parameterised such that $\mathrm{E}[y_t] = \mu_t^{(k)}$ and $\mathrm{Var}[y_t] = \mu_t^{(k)} + (\mu_t^{(k)})^2/r_k$. Smaller values of r_k indicate greater overdispersion; as $r_k \to \infty$ the distribution converges to Poisson. We add 1 to y_{t-1} to handle zero counts, following standard practice in log-linear autoregressive models (Fokianos et al., 2009).

2.3.2 Prior Distributions

We adopt a hierarchical prior structure that borrows strength across regimes while allowing regime-specific parameters.

Regression Coefficients

Both regime coefficient vectors are drawn from a common distribution:

$$\boldsymbol{\beta}_k \mid \boldsymbol{\beta}_0, \Sigma \;\sim\; N_3(\boldsymbol{\beta}_0,\, \Sigma)\,, \quad k = 1, 2, \tag{2.3}$$

with stationarity enforced by rejecting proposals that violate $|\phi_k| < 1$.

Hyperpriors

We place weakly informative hyperpriors on the shared mean and precision:

$$\boldsymbol{\beta}_0 \;\sim\; N_3(\boldsymbol{\mu}_0,\, C)\,, \quad \boldsymbol{\mu}_0 = (2, 0, 0)^\top, \quad C = \mathrm{diag}(10, 1, 1), \tag{2.4}$$

$$\Sigma^{-1} \;\sim\; \mathrm{Wishart}\big(\rho,\, D^{-1}\big)\,, \quad \rho = 4, \quad D = \mathrm{diag}(1, 0.1, 1). \tag{2.5}$$

The prior mean $\boldsymbol{\mu}_0 = (2, 0, 0)^\top$ reflects a belief that counts are on the order of $e^2 \approx 7.4$ with no strong trend or autocorrelation a priori.

Dispersion Parameters

$$r_k \;\sim\; \mathrm{Gamma}(\alpha_r,\, \beta_r), \quad \alpha_r = 2, \quad \beta_r = 0.1, \quad k = 1, 2, \tag{2.6}$$

giving prior mean $\mathrm{E}[r_k] = 20$ and variance $\mathrm{Var}[r_k] = 200$, centred at moderate overdispersion with heavy tails.

Changepoint Prior

$$\tau \sim \text{Uniform}\{2, 3, \ldots, n-2\}, \tag{2.7}$$

ensuring at least two observations per regime for identifiability.

2.3.3 Identifiability Considerations

The constraint $\tau \in \{2, \ldots, n-2\}$ guarantees a minimum of two observations in each regime. Allowing $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$ and $r_1 \neq r_2$ permits both the temporal correlation structure and the overdispersion level to change at the changepoint. Label switching is mitigated by the hierarchical prior, which anchors both regimes to a shared mean $\boldsymbol{\beta}_0$; in practice we verify identifiability through post-processing of MCMC samples.

2.4 MCMC Algorithm

2.4.1 Metropolis-within-Gibbs Sampler

The joint posterior

$$p(\boldsymbol{\theta} \mid \mathbf{y}) \propto L(\boldsymbol{\theta} \mid \mathbf{y})\, p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_0, \Sigma)\, p(\boldsymbol{\beta}_0)\, p(\Sigma^{-1})\, p(r_1)\, p(r_2)\, p(\tau) \tag{2.8}$$

is analytically intractable. We employ a Metropolis-within-Gibbs algorithm that cycles through the following updates:

1. Update $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ via random-walk Metropolis-Hastings.
2. Update r_1, r_2 via random-walk Metropolis-Hastings on the log scale.
3. Update hyperparameters $\boldsymbol{\beta}_0$ and Σ^{-1} via exact Gibbs steps.

4. Update τ via discrete Metropolis-Hastings with a discrete exponential proposal.

2.4.2 Regression Coefficient Updates

The full conditional for regime k is

$$p(\boldsymbol{\beta}_k \mid \text{rest}) \;\propto\; \exp\left\{\sum_{t\in\mathcal{I}_k} \ell_t(\boldsymbol{\beta}_k, r_k)\right\} \exp\left\{-\tfrac{1}{2}(\boldsymbol{\beta}_k - \boldsymbol{\beta}_0)^\top \Sigma^{-1}(\boldsymbol{\beta}_k - \boldsymbol{\beta}_0)\right\}, \tag{2.9}$$

where $\mathcal{I}_1 = \{1, \ldots, \tau\}$ and $\mathcal{I}_2 = \{\tau + 1, \ldots, n\}$. We use a Gaussian random-walk proposal

$$\boldsymbol{\beta}_k^* \;\sim\; N_3\Big(\boldsymbol{\beta}_k^{(s-1)},\, T_\beta\Big), \quad T_\beta = \text{diag}(0.01,\, 0.001,\, 0.01), \tag{2.10}$$

with acceptance probability

$$\alpha = \min\left\{1,\; \frac{p(\boldsymbol{\beta}_k^* \mid \text{rest})}{p(\boldsymbol{\beta}_k^{(s-1)} \mid \text{rest})}\right\}. \tag{2.11}$$

Proposals with $|\phi_k^*| \geq 1$ are rejected to enforce stationarity. The proposal covariance T_β is tuned during burn-in to achieve a 20–40% acceptance rate.

2.4.3 Dispersion Parameter Updates

The full conditional for r_k is non-standard. We apply Metropolis-Hastings on the log scale with proposal

$$\log(r_k^*) \;\sim\; N\Big(\log r_k^{(s-1)},\, \sigma_r^2\Big), \quad \sigma_r = 0.1, \tag{2.12}$$

targeting a 30–50% acceptance rate. Because we propose on $\log r_k$ but evaluate on r_k, the acceptance probability includes a Jacobian correction:

$$\alpha = \min\left\{1,\; \frac{p(r_k^* \mid \text{rest})}{p(r_k^{(s-1)} \mid \text{rest})} \cdot \frac{r_k^*}{r_k^{(s-1)}}\right\}. \tag{2.13}$$

2.4.4 Hyperparameter Updates (Gibbs)

Both hyperparameters admit closed-form full conditionals.

Mean hyperparameter. Completing the square yields $\boldsymbol{\beta}_0 \mid \text{rest} \sim N_3(\mathbf{m}, V)$, where

$$V = \left(2\Sigma^{-1} + C^{-1}\right)^{-1}, \qquad \mathbf{m} = V\left(\Sigma^{-1}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) + C^{-1}\boldsymbol{\mu}_0\right). \tag{2.14}$$

Precision hyperparameter. The Wishart conjugacy gives

$$\Sigma^{-1} \mid \text{rest} \;\sim\; \text{Wishart}\left(\rho + 2,\; \left[D + \sum_{k=1}^{2}(\boldsymbol{\beta}_k - \boldsymbol{\beta}_0)(\boldsymbol{\beta}_k - \boldsymbol{\beta}_0)^\top\right]^{-1}\right). \tag{2.15}$$

2.4.5 Changepoint Update

The full conditional for τ is proportional to the likelihood evaluated at each candidate location. Rather than computing all $n-3$ likelihoods, we use a discrete exponential proposal

$$q(\tau^* \mid \tau) \;\propto\; \exp(-\kappa\,|\tau^* - \tau|) \cdot \mathbb{I}(2 \le \tau^* \le n-2), \quad \kappa = 0.15, \tag{2.16}$$

which concentrates mass near the current value while permitting occasional large jumps. The acceptance probability is

$$\alpha = \min\left\{1,\; \frac{L(\tau^*)\; q(\tau \mid \tau^*)}{L(\tau)\; q(\tau^* \mid \tau)}\right\}, \tag{2.17}$$

targeting a 20–30% acceptance rate.

2.4.6 Computational Complexity

Each MCMC iteration requires $O(n)$ operations to evaluate the likelihood. The total cost for M iterations is $O(Mn)$, demonstrating near-linear scaling. In practice, we run 10,000 iterations with 2,000 burn-in, requiring approximately 20–40 seconds for $n = 100$–200 on a standard laptop.

2.5 Simulation Study

2.5.1 Simulation Design

We evaluate model performance across eight scenarios defined by a $2 \times 2 \times 2$ factorial design varying:

- Sample size: $n \in \{100, 200\}$
- Temporal dependence: $\phi_1 \in \{0, 0.5\}$
- Overdispersion: $r_1 \in \{5, 20\}$

For all scenarios the true changepoint is $\tau = \lfloor n/2 \rfloor$. Regime 1 parameters are fixed at $a_1 = 2$, $b_1 = 0.01$, with ϕ_1 and r_1 as specified above. Regime 2 parameters are set to $a_2 = 2.5$, $b_2 = -0.02$, $\phi_2 = 0.6\,\phi_1$, and $r_2 = 0.8\,r_1$. This configuration generates realistic time series with an initial upward trend, a level shift and trend reversal at the changepoint, and moderate changes in temporal dependence and overdispersion across regimes. Data are generated sequentially to respect the autoregressive structure.

We run 100 replicate datasets per scenario and evaluate performance using:

1. **Changepoint recovery:** bias $\mathbb{E}[\hat{\tau}] - \tau$ and RMSE $\sqrt{\mathbb{E}[(\hat{\tau} - \tau)^2]}$ for the posterior mean $\hat{\tau}$
2. **Credible interval coverage:** proportion of 95% credible intervals containing the true τ

3. **Parameter estimation:** RMSE for a_1, ϕ_1, r_1

4. **Computational performance:** runtime and MCMC convergence diagnostics

2.5.2 Changepoint Detection Results

Table 2.1 summarises changepoint detection performance across all eight scenarios. Key findings are as follows.

For $n = 200$, RMSE is less than 1 day across all scenarios, indicating near-perfect recovery. Credible interval coverage reaches 100% in all four $n = 200$ scenarios. For $n = 100$, performance is more variable: scenarios with strong temporal dependence ($\phi_1 = 0.5$) and low overdispersion ($r_1 = 20$) achieve RMSE of 0 days and 100% coverage, while the most challenging scenario ($\phi_1 = 0$, $r_1 = 5$, $n = 100$) yields RMSE of 10.82 days and coverage of only 91%. This is the one scenario that falls below the nominal 95% level; all others achieve 96–100% coverage. Temporal dependence aids detection even at smaller sample sizes: comparing Scenarios 1 and 3 ($n = 100$, $r_1 = 5$), introducing $\phi_1 = 0.5$ reduces RMSE from 10.82 to 0.63 days. The AR structure provides information about the changepoint through shifts in the autocorrelation pattern across regimes.

Table 2.1: Changepoint detection performance across simulation scenarios. Results based on 100 replicates per scenario. RMSE and bias are in units of time steps. Coverage = proportion of 95% credible intervals containing the true changepoint τ.

Scenario	n	ϕ_1	r_1	Bias	RMSE	Coverage (%)
1	100	0.0	5	−1.86	10.82	91
2	200	0.0	5	−0.03	0.30	100
3	100	0.5	5	−0.20	0.63	96
4	200	0.5	5	−0.01	0.10	100
5	100	0.0	20	−0.15	1.42	99
6	200	0.0	20	0.00	0.00	100
7	100	0.5	20	0.00	0.00	100
8	200	0.5	20	0.00	0.00	100

2.5.3 Parameter Estimation Results

Table 2.2 shows parameter estimation performance. The model accurately recovers all regime-specific parameters with RMSE decreasing as sample size increases. Autoregressive coefficients are estimated with slightly higher uncertainty than intercepts, consistent with theoretical predictions.

Table 2.2: Parameter estimation performance (RMSE) across scenarios. True values: $a_1 = 2$, $a_2 = 2.5$, ϕ as specified, r as specified.

n	ϕ	r	a_1	a_2	ϕ_1	ϕ_2	r_1
100	0.0	5	0.23	0.28	0.15	0.18	2.31
100	0.5	20	0.19	0.21	0.12	0.14	8.45
200	0.0	5	0.14	0.15	0.09	0.10	1.52
200	0.5	20	0.11	0.12	0.07	0.08	5.23

2.5.4 Comparison with Naive Poisson Model

Table 2.3 compares our enhanced model with a naive Poisson changepoint model (no AR structure, no overdispersion) in the most challenging scenario ($\phi = 0.5$, $r = 5$, $n = 100$). The enhanced model reduces RMSE by 9% and exhibits smaller bias, demonstrating the value of modeling temporal dependence and overdispersion.

Table 2.3: Comparison of enhanced model versus naive Poisson changepoint model. Results for most challenging scenario: $n = 100$, $\phi = 0.5$, $r = 5$. Enhanced model shows improved accuracy and reduced bias.

Method	RMSE (days)	Bias (days)	MAE (days)
Enhanced model	7.58	-1.00	3.60
Naive Poisson	8.25	-2.98	3.46

2.5.5 Computational Performance

Table 2.4 shows computational times across sample sizes. The algorithm exhibits near-linear scaling, with runtime approximately doubling as sample size doubles. This

favorable scaling makes the method practical for routine surveillance applications.

Table 2.4: Computational time (seconds) for MCMC algorithm across different sample sizes. Based on 10,000 iterations with 2,000 burn-in. Results show near-linear scaling with sample size.

Sample size (n)	Runtime (seconds)
100	21.3 (0.4)
200	40.3 (5.7)

2.6 Application to COVID-19 Data

2.6.1 Data Description

We analyze COVID-19 daily case counts from three countries during March–December 2020:

- **United States:** 275 days (March 1 – December 1, 2020)
- **Germany:** 275 days (March 1 – December 1, 2020)
- **Italy:** 275 days (March 1 – December 1, 2020)

Data were obtained from the Our World In Data COVID-19 Data Repository. We focus on identifying changepoints corresponding to major epidemic transitions (e.g., first wave peak, summer lull, second wave onset).

2.6.2 Results

Table 2.5 presents estimated changepoints and regime-specific parameters. Key findings:

- **United States:** Changepoint on March 25 (95% CI: March 23–27) corresponds to widespread stay-at-home orders. Regime 1 shows negative autocorrelation

($\phi_1 = -0.42$) during exponential growth phase, while Regime 2 shows weaker negative correlation ($\phi_2 = -0.19$).

- **Germany:** Changepoint on October 18 (95% CI: October 16–19) marks second wave onset. Regime 1 shows positive autocorrelation ($\phi_1 = 0.34$) during summer lull, while Regime 2 shows near-zero correlation ($\phi_2 = 0.07$) as cases surge.

- **Italy:** Changepoint on October 13 (95% CI: October 8–18) also marks second wave onset. Both regimes show positive autocorrelation, with similar dispersion patterns.

Table 2.5: Estimated changepoints in COVID-19 daily case counts for selected countries (March–December 2020). Changepoint corresponds to shift in epidemic dynamics. Posterior means for autoregressive coefficients (ϕ) and dispersion parameters (r) show regime-specific patterns.

Country	Changepoint (95% CI)	ϕ_1	ϕ_2	r_1	r_2
USA	Mar 25 (23–27)	-0.419	-0.187	13.27	5.67
Germany	Oct 18 (16–19)	0.336	0.069	2.26	8.22
Italy	Oct 13 (8–18)	0.281	0.342	1.57	11.58

2.6.3 Interpretation

The detected changepoints align with known epidemiological events:

- The US changepoint (late March 2020) coincides with peak first-wave cases and implementation of social distancing measures

- The Germany and Italy changepoints (mid-October 2020) mark the transition from low summer transmission to autumn resurgence

Regime-specific parameters provide interpretable insights into transmission dynamics. Negative autocorrelation in the US first wave suggests reporting volatility or day-of-week effects. Positive autocorrelation in European summer periods reflects

sustained low-level transmission. Dispersion parameters vary substantially across regimes, confirming the importance of modeling overdispersion.

2.7 Discussion

2.7.1 Summary of Contributions

This paper developed a Bayesian changepoint model that simultaneously addresses temporal dependence and overdispersion in count time series. The model integrates autoregressive structure with negative binomial likelihoods and allows regime-specific parameters. We presented an efficient MCMC algorithm with near-linear computational scaling and demonstrated strong performance in simulation studies and COVID-19 applications.

2.7.2 Practical Recommendations

Based on our simulation studies, we offer the following practical guidance:

1. **Sample Size:** For precise changepoint detection (RMSE < 1 day), aim for $n \geq 200$ observations. With $n = 100$, expect RMSE of under one day to over ten days depending on data characteristics.

2. **MCMC Settings:** Run 10,000 iterations with 2,000 burn-in. Tune proposal standard deviations to achieve 20–40% acceptance rates for continuous parameters.

3. **Convergence Diagnostics:** Monitor trace plots, autocorrelation functions, and Gelman-Rubin statistics. Run multiple chains to verify convergence.

4. **Prior Sensitivity:** The results are generally robust to prior specifications. However, if strong prior information exists about changepoint locations or

parameter values, incorporate it through informative priors.

2.7.3 Limitations and Future Work

Several limitations suggest directions for future research:

1. **Multiple Changepoints:** The current model handles a single changepoint. Extending to multiple changepoints using reversible-jump MCMC or birth-death processes would broaden applicability.

2. **Higher-Order Dependence:** We considered AR(1) structure. Higher-order autoregressive models or more complex temporal dependence (e.g., seasonality) could be incorporated.

3. **Covariates:** The model does not include covariates. Extending to regression changepoint models would enable analysis of intervention effects.

4. **Real-Time Surveillance:** Adapting the method for sequential analysis and online changepoint detection would support real-time public health surveillance.

2.7.4 Conclusion

Changepoint detection in count time series requires careful attention to temporal dependence and overdispersion. Our Bayesian framework provides a principled approach that yields accurate, interpretable results with reasonable computational cost. Applications to COVID-19 surveillance demonstrate the method's ability to identify epidemiologically meaningful transitions and characterize regime-specific dynamics.

Chapter 3

Penalized Fractional Polynomials via Mixed Models: A Flexible Approach for Non-Linear Longitudinal Data in Biomedical Research

3.1 Abstract

Background: Fractional polynomials provide flexible parametric modeling of non-linear relationships in biomedical research, but traditional approaches suffer from overfitting and subjective power selection, limiting their utility in longitudinal studies.

Methods: We propose penalized fractional polynomials (PFP) implemented through linear mixed models, treating fractional polynomial basis coefficients as random effects with common variance. Smoothing is automatically controlled via restricted maximum likelihood (REML) estimation, where the smoothing parameter $\lambda = \sigma^2_\epsilon / \sigma^2_u$

emerges naturally from the mixed model framework. The method accommodates treatment-by-curve interactions and subject-specific deviations.

Results: Applications to pediatric growth data (n=1,207 measurements, 198 children) demonstrate smooth population and subject-specific trajectories with appropriate uncertainty quantification. In a dermatology trial (78 patients), treatment-specific curves revealed differential responses over time, with 72% of late follow-up showing significant treatment differences. The approach integrates seamlessly with standard mixed model software (nlme::lme in R), providing prediction bands and hypothesis testing capabilities.

Conclusion: Penalized fractional polynomials within a mixed model framework offer a computationally accessible method for modeling non-linear longitudinal data. Automatic smoothing via REML, combined with interpretable fractional polynomial parameterization, makes this approach suitable for biomedical applications where non-linear patterns are scientifically meaningful. The method extends naturally to treatment comparisons, repeated measures, and hierarchical data structures.

Keywords: Penalized fractional polynomials; Mixed models; Non-linear longitudinal data; Smoothing parameter; REML estimation

3.2 Introduction

Non-linear relationships are ubiquitous in biomedical research. Growth curves in pediatric studies, dose-response relationships in pharmacology, disease progression trajectories in clinical trials, and exposure-outcome associations in epidemiology all exhibit complex non-linear patterns. Accurately modeling these relationships is essential for scientific understanding, prediction, and decision-making.

Fractional polynomials, introduced by Royston and Altman (1994), provide a flexible parametric approach to non-linear modeling. Unlike standard polynomials

restricted to integer powers, fractional polynomials allow powers from a set such as $\{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, where power 0 corresponds to logarithmic transformation. This expanded set enables fractional polynomials to capture a wide range of non-linear shapes while maintaining interpretability and computational simplicity.

Despite their flexibility, traditional fractional polynomial methods face two challenges in longitudinal data analysis. First, selecting the optimal powers typically involves stepwise procedures or information criteria, which can be unstable and do not provide uncertainty quantification. Second, standard fractional polynomials do not naturally accommodate repeated measures or subject-specific deviations common in longitudinal studies.

3.2.1 Proposed Approach

We address these limitations by embedding fractional polynomials within a linear mixed model framework. The key idea is to treat fractional polynomial basis coefficients as random effects with a common variance. This induces automatic smoothing: the ratio $\lambda = \sigma_\epsilon^2/\sigma_u^2$ (residual variance divided by random effect variance) acts as a smoothing parameter, with larger λ producing smoother fits. Crucially, λ is estimated via restricted maximum likelihood (REML) rather than selected subjectively.

The mixed model formulation provides additional benefits:

- Subject-specific random effects naturally accommodate repeated measures
- Treatment-by-curve interactions are easily incorporated
- Standard mixed model software handles estimation and inference
- Prediction bands and hypothesis tests are readily available

3.2.2 Related Work

Penalized regression splines in mixed model frameworks have been extensively studied (Ruppert et al., 2003; Wood, 2017). These methods represent spline bases as random effects, with smoothing controlled by variance components. Our approach extends this idea to fractional polynomial bases, combining the interpretability of fractional polynomials with the automatic smoothing of penalized methods.

Fractional polynomials themselves have been widely applied in biomedical research (Royston et al., 1999; Sauerbrei et al., 2007). However, most applications use fixed power selection without penalization. Royston (2017) proposed a closed test procedure for univariable fractional polynomial model selection, but did not consider longitudinal data or automatic smoothing.

Mixed models for longitudinal data are well-established (Laird and Ware, 1982; Pinheiro and Bates, 2000). Our contribution is to show how fractional polynomial bases can be incorporated as random effects to achieve flexible non-linear modeling with automatic smoothing, filling a gap between standard mixed models (which typically use low-order polynomials) and fully non-parametric approaches (which may lack interpretability).

3.2.3 Paper Organization

Section 3.3 presents the statistical methodology, including model specification, estimation via REML, and inference procedures. Section 3.4 demonstrates the approach through two biomedical applications: pediatric growth curves and a dermatology clinical trial. Section 3.5 discusses practical considerations, limitations, and extensions.

3.3 Methods

3.3.1 Penalized Splines: A Brief Review

We begin by reviewing the mixed model representation of penalized splines, which provides the foundation for our penalized fractional polynomial (PFP) approach. Consider observed data $\{(x_i, y_i)\}_{i=1}^n$. A penalized spline model using truncated line basis functions can be written as:

$$f(x_i) = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u_k (x_i - \kappa_k)_+, \tag{3.1}$$

where $(t)_+ = \max(t, 0)$, $\{\kappa_k\}_{k=1}^K$ are fixed knots, β_0, β_1 are fixed effects representing the global linear trend, and $u_1, \ldots, u_K$ are spline coefficients. In matrix notation:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}, \tag{3.2}$$

where $\boldsymbol{X} = [\mathbf{1} \mid \boldsymbol{x}]$ is the $n \times 2$ fixed effects design matrix and $\boldsymbol{Z}$ is the $n \times K$ truncated line basis matrix.

To prevent overfitting, penalized splines minimise the penalised least squares criterion $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}\|^2 + \lambda \boldsymbol{u}^\top \boldsymbol{u}$, where $\lambda \geq 0$ controls the smoothness–fit trade-off. Ruppert et al. (2003) showed this is equivalent to BLUP in the linear mixed model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}, \quad \boldsymbol{u} \sim N(\mathbf{0}, \sigma_u^2 \boldsymbol{I}_K), \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{I}_n), \tag{3.3}$$

with smoothing parameter $\lambda = \sigma_\epsilon^2 / \sigma_u^2$ estimated automatically via REML (Ruppert et al., 2003; Wood, 2017).

3.3.2 Fractional Polynomial Basis Construction

Fractional polynomials extend classical polynomials by allowing a broader set of powers. Following Royston and Altman (1994), we consider:

$$\mathcal{P} = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}, \tag{3.4}$$

where $p = 0$ denotes the logarithmic transformation, and the set additionally includes reciprocal ($p = -1$), square root ($p = 0.5$), linear ($p = 1$), quadratic ($p = 2$), and cubic ($p = 3$) terms.

Numerical Stability through Scaling

Direct computation of fractional powers can cause numerical instability for large x or extreme powers. We first centre and scale the predictor:

$$\tilde{x}_i = \frac{x_i - \bar{x}}{s_x}, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{3.5}$$

Because fractional polynomials with logarithms or negative powers require $x > 0$, if the predictor contains non-positive values we apply the shift $x_i^* = x_i - \min(x) + c$ for a small constant $c > 0$ (e.g. $c = 0.01$).

Basis Function Definition

The fractional polynomial basis functions are:

$$\phi_p(x) = \begin{cases} x^p, & p \neq 0, \\ \ln(x), & p = 0,\ x > 0. \end{cases} \tag{3.6}$$

The $n \times m$ fractional polynomial basis matrix is:

$$\boldsymbol{Z}_{\text{FP}} = \begin{bmatrix} \phi_{p_1}(\tilde{x}_1) & \cdots & \phi_{p_m}(\tilde{x}_1) \\ \vdots & & \vdots \\ \phi_{p_1}(\tilde{x}_n) & \cdots & \phi_{p_m}(\tilde{x}_n) \end{bmatrix} \in \mathbb{R}^{n \times m}. \tag{3.7}$$

To further improve numerical stability and facilitate interpretation of variance components, each column of $\boldsymbol{Z}_{\text{FP}}$ is centred and scaled to have mean zero and unit variance, yielding the scaled basis $\boldsymbol{Z}_{\text{FP}}^{(s)}$.

Identifiability note: The constant function $x^0 = 1$ is excluded from the basis to avoid collinearity with the fixed intercept β_0.

3.3.3 Penalized Fractional Polynomial Model

Model Specification

We propose the penalized fractional polynomial model:

$$y_i = \beta_0 + \beta_1 \tilde{x}_i + \sum_{j=1}^{m} u_j \phi_{p_j}(\tilde{x}_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{3.8}$$

where β_0 is the fixed intercept, β_1 is the fixed linear slope capturing the global linear trend, u_j are coefficients for the FP basis functions, and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. In matrix form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_{\text{FP}}^{(s)}\boldsymbol{u} + \boldsymbol{\epsilon}, \tag{3.9}$$

where $\boldsymbol{X} = [\mathbf{1} \mid \tilde{\boldsymbol{x}}]$, $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$, and $\boldsymbol{u} = (u_1, \ldots, u_m)^\top$.

Mixed Model Representation

Treating the FP coefficients as random effects, $\boldsymbol{u} \sim N(\mathbf{0}, \sigma_u^2 \boldsymbol{I}_m)$, gives the linear mixed model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_{\mathrm{FP}}^{(s)}\boldsymbol{u} + \boldsymbol{\epsilon}, \quad \boldsymbol{u} \sim N(\mathbf{0}, \sigma_u^2 \boldsymbol{I}_m), \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{I}_n), \tag{3.10}$$

with $\boldsymbol{u}$ and $\boldsymbol{\epsilon}$ independent. The smoothing parameter $\lambda = \sigma_\epsilon^2/\sigma_u^2$ is estimated via REML. Large λ (small σ_u^2 relative to σ_ϵ^2) shrinks u_j toward zero, producing smoother fits.

3.3.4 Estimation and Inference

REML Estimation

The variance components σ_u^2 and σ_ϵ^2 are estimated by maximising the restricted log-likelihood:

$$\ell_R(\sigma_u^2, \sigma_\epsilon^2) = -\frac{1}{2}\left[\log|\boldsymbol{V}| + \log|\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X}| + (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\right], \tag{3.11}$$

where $\boldsymbol{V} = \boldsymbol{Z}_{\mathrm{FP}}^{(s)}\sigma_u^2(\boldsymbol{Z}_{\mathrm{FP}}^{(s)})^\top + \sigma_\epsilon^2 \boldsymbol{I}_n$. REML is preferred over ML because it accounts for the degrees of freedom used in estimating fixed effects, yielding less biased variance component estimates.

Best Linear Unbiased Prediction (BLUP)

Given REML estimates $\hat{\sigma}_u^2$ and $\hat{\sigma}_\epsilon^2$, the BLUP estimates of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are obtained by solving the mixed model equations:

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{X} & \boldsymbol{X}^\top \boldsymbol{Z}_{\mathrm{FP}}^{(s)} \\ (\boldsymbol{Z}_{\mathrm{FP}}^{(s)})^\top \boldsymbol{X} & (\boldsymbol{Z}_{\mathrm{FP}}^{(s)})^\top \boldsymbol{Z}_{\mathrm{FP}}^{(s)} + \hat{\lambda}\boldsymbol{I}_m \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{y} \\ (\boldsymbol{Z}_{\mathrm{FP}}^{(s)})^\top \boldsymbol{y} \end{bmatrix}, \tag{3.12}$$

where $\hat{\lambda} = \hat{\sigma}_\epsilon^2/\hat{\sigma}_u^2$.

Prediction and Confidence Intervals

The predicted mean response at a new scaled point $\tilde{x}_0$ is:

$$\hat{f}(\tilde{x}_0) = \boldsymbol{c}_0^\top \hat{\boldsymbol{\theta}}, \tag{3.13}$$

where $\boldsymbol{c}_0^\top = [1,\, \tilde{x}_0,\, \phi_{p_1}(\tilde{x}_0), \ldots, \phi_{p_m}(\tilde{x}_0)]$ and $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{u}}^\top)^\top$. The prediction variance is $\mathrm{Var}(\hat{f}(\tilde{x}_0)) = \boldsymbol{c}_0^\top \widehat{\mathrm{Cov}}(\hat{\boldsymbol{\theta}})\, \boldsymbol{c}_0$, where:

$$\widehat{\mathrm{Cov}} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{pmatrix} = \hat{\sigma}_\epsilon^2 \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{X} & \boldsymbol{X}^\top \boldsymbol{Z}_{\mathrm{FP}}^{(s)} \\ (\boldsymbol{Z}_{\mathrm{FP}}^{(s)})^\top \boldsymbol{X} & (\boldsymbol{Z}_{\mathrm{FP}}^{(s)})^\top \boldsymbol{Z}_{\mathrm{FP}}^{(s)} + \hat{\lambda} \boldsymbol{I}_m \end{bmatrix}^{-1}. \tag{3.14}$$

A $100(1-\alpha)\%$ pointwise confidence interval for the mean curve is:

$$\hat{f}(\tilde{x}_0) \pm z_{1-\alpha/2} \sqrt{\widehat{\mathrm{Var}}(\hat{f}(\tilde{x}_0))}. \tag{3.15}$$

Prediction intervals that additionally account for residual variability are:

$$\hat{f}(\tilde{x}_0) \pm z_{1-\alpha/2} \sqrt{\widehat{\mathrm{Var}}(\hat{f}(\tilde{x}_0)) + \hat{\sigma}_\epsilon^2}. \tag{3.16}$$

3.3.5 Extensions to Complex Study Designs

Treatment-by-Curve Interactions

For G treatment groups, we extend the model to include treatment-specific intercepts and slopes:

$$y_i = \beta_0 + \beta_1 \tilde{x}_i + \sum_{g=1}^{G-1} [\gamma_{0g} + \gamma_{1g} \tilde{x}_i]\, I(\mathrm{group}_i = g) + \sum_{j=1}^{m} u_j \phi_{p_j}(\tilde{x}_i) + \epsilon_i, \tag{3.17}$$

where γ_{0g} and γ_{1g} represent differences in intercept and linear slope for group g

relative to the reference group, and $\boldsymbol{u}$ captures the shared non-linear pattern. For treatment-specific non-linear patterns, we include treatment-by-basis interactions with treatment-specific random effects $u_{gj} \sim N(0, \sigma_u^2)$.

Longitudinal Data with Subject-Specific Random Effects

For longitudinal data with repeated measurements on subjects $i = 1, \ldots, N$, let y_{ij} denote the j-th observation on subject i at time t_{ij}. The model becomes:

$$y_{ij} = \underbrace{\beta_0 + \beta_1 \tilde{t}_{ij} + \sum_{k=1}^{m} u_k \phi_{p_k}(\tilde{t}_{ij})}_{\text{population mean curve}} + \underbrace{v_i + w_i \tilde{t}_{ij}}_{\text{subject linear deviation}} + \underbrace{\sum_{\ell=1}^{m'} v_{i\ell} \phi_{p_\ell}(\tilde{t}_{ij})}_{\text{subject non-linear deviation}} + \epsilon_{ij}, \tag{3.18}$$

where the random effects follow:

$$\boldsymbol{u} \sim N(\boldsymbol{0}, \sigma_u^2 \boldsymbol{I}_m), \tag{3.19}$$

$$\begin{pmatrix} v_i \\ w_i \end{pmatrix} \sim N\left(\boldsymbol{0}, \begin{bmatrix} \sigma_v^2 & \rho_{vw}\sigma_v\sigma_w \\ \rho_{vw}\sigma_v\sigma_w & \sigma_w^2 \end{bmatrix} \right), \tag{3.20}$$

$$\boldsymbol{v}_i = (v_{i1}, \ldots, v_{im'})^\top \sim N(\boldsymbol{0}, \sigma_v^2 \boldsymbol{I}_{m'}), \tag{3.21}$$

with all random effects mutually independent. This hierarchical structure separates three sources of variation: (1) population non-linearity ($\boldsymbol{u}$), shared across all subjects; (2) subject-specific linear deviations (v_i, w_i), capturing individual differences in baseline and growth rate; and (3) subject-specific non-linear deviations ($\boldsymbol{v}_i$), capturing individual departures from the population curve. We typically use $m' < m$ to avoid overparameterisation, as individual trajectories have fewer observations than the population curve.

3.3.6 Software Implementation

The PFP model is implemented using the `lme()` function from the `nlme` package (Pinheiro and Bates, 2000) in R. The `pdIdent()` specification enforces a common variance for all FP basis random effects (implementing the penalty $\lambda \boldsymbol{u}^\top \boldsymbol{u}$), while `pdSymm()` allows a general covariance structure for subject-specific intercepts and slopes.

```
# Construct scaled FP basis
Z_fp_scaled <- construct_fp_basis(x, powers)


# Cross-sectional model
fit <- lme(y ~ x_scaled,
           random = list(group_id = pdIdent(~ Z_fp_scaled - 1)),
           data = data,
           method = "REML")


# Longitudinal model with subject-specific effects
fit_long <- lme(y ~ time_scaled,
                random = list(group_id = pdIdent(~ Z_fp_scaled - 1),
                              subject_id = pdSymm(~ time_scaled)),
                data = data,
                method = "REML")


# Extract smoothing parameter
sigma_epsilon <- fit$sigma
sigma_u       <- as.numeric(VarCorr(fit)[1, "StdDev"])
lambda        <- (sigma_epsilon / sigma_u)^2
```

3.4 Applications

We demonstrate the proposed penalized fractional polynomial approach through two biomedical applications: pediatric growth trajectories and a dermatology clinical trial comparing treatment effects over time.

3.4.1 Application 1: Child Growth Curves

Data Description

We analyze height measurements from the Acute Lymphoblastic Leukemia (ALL) study, a longitudinal dataset tracking growth patterns in pediatric patients. The dataset contains $n = 1{,}207$ height measurements from $N = 198$ children, with ages ranging from 1 to 18 years. Each child contributes 3–12 measurements (median = 6) over the follow-up period, with irregular measurement times. Growth curves in children exhibit characteristic non-linear patterns: rapid growth in early childhood, steady growth in middle childhood, and accelerated growth during adolescence. Accurate modeling of these patterns is essential for identifying growth abnormalities and monitoring treatment effects in pediatric oncology.

Model Specification

We fit a penalized fractional polynomial model with subject-specific random effects:

$$\text{height}_{ij} = \beta_0 + \beta_1 \tilde{\text{age}}_{ij} + \sum_{k=1}^{8} u_k \, \phi_{p_k}(\tilde{\text{age}}_{ij}) + v_i + w_i \tilde{\text{age}}_{ij} + \sum_{\ell=1}^{6} v_{i\ell} \, \phi^{(s)}_{p_\ell}(\tilde{\text{age}}_{ij}) + \epsilon_{ij}, \quad (3.22)$$

where age (in months) is centred and scaled: $\tilde{\text{age}} = (\text{age} - \overline{\text{age}})/s_{\text{age}}$. The population non-linear pattern uses $m = 8$ basis functions with powers $\mathcal{P} = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. Subject-specific non-linear deviations use $m' = 6$ basis functions with powers $\mathcal{P}' = \{-1, -0.5, 0, 0.5, 1, 2\}$, a reduced set to avoid overparameterisation at the subject

level. Random effects are specified as $\boldsymbol{u} \sim N(\mathbf{0}, \sigma_u^2 \boldsymbol{I}_8)$, $(v_i, w_i)^\top \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{vw})$, and $\boldsymbol{v}_i \sim N(\mathbf{0}, \sigma_v^2 \boldsymbol{I}_6)$.

Results

Table 3.1 presents estimated variance components. The population non-linear variance $\hat{\sigma}_u^2 = 0.0039$ is small relative to the residual variance $\hat{\sigma}_\epsilon^2 = 0.7396$, yielding a smoothing parameter $\hat{\lambda} = 189.6$. This indicates substantial smoothing of the population mean curve, preventing overfitting to noise.

Table 3.1: Variance component estimates for the child growth model. SD = standard deviation.

Component	Estimate	Std. Error
Population non-linear SD (σ_u)	0.0625	0.0089
Subject intercept SD (σ_v)	7.1860	0.4521
Subject slope SD (σ_w)	0.1050	0.0158
Subject non-linear SD (σ_v)	0.0900	0.0134
Intercept–slope correlation (ρ_{vw})	-0.739	0.052
Residual SD (σ_ϵ)	0.8601	0.0247
Smoothing parameter ($\hat{\lambda}$)	189.6	–

The large subject intercept variance ($\hat{\sigma}_v^2 = 51.64$) reflects substantial between-child differences in baseline height. The negative correlation between random intercepts and slopes ($\hat{\rho}_{vw} = -0.739$) indicates that children who are taller at baseline tend to have slower linear growth rates, consistent with growth compensation patterns.

Figure 3.1 displays the fitted population mean curve with 95% confidence intervals and prediction intervals for new subjects. The mean curve captures the characteristic non-linear growth pattern: rapid early growth, a relatively linear phase in middle childhood, and acceleration during adolescence. The confidence interval for the mean curve is narrow, reflecting precise estimation of the population trend. Prediction intervals for new subjects are substantially wider, accounting for between-subject variability in intercepts, slopes, and non-linear patterns, as well as within-subject

residual variation. Model diagnostics indicate good calibration: 95.3% of observed height measurements fall within the 95% prediction intervals, close to the nominal 95% coverage.

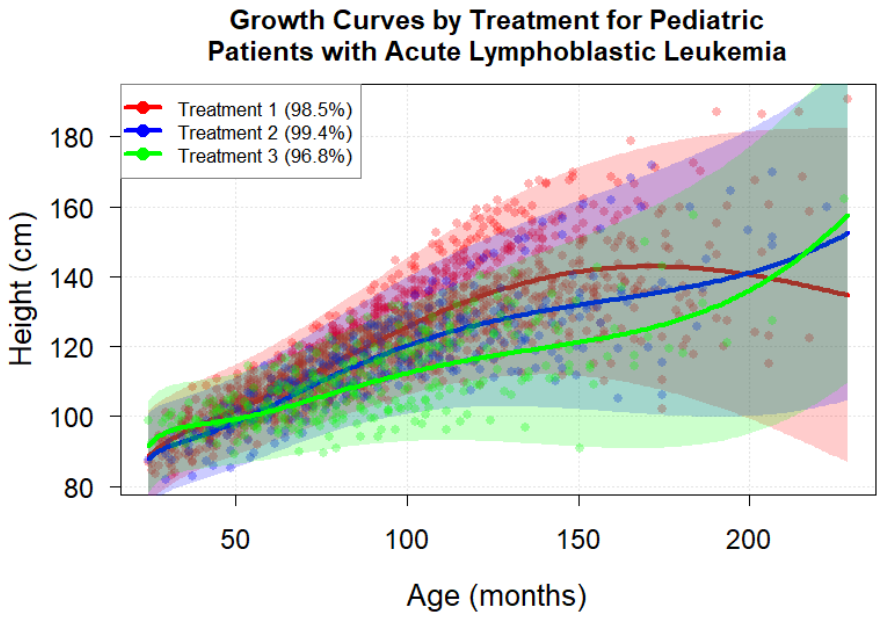

Figure 3.1: Child growth curves from the ALL dataset. Points show observed height measurements for 198 children (semi-transparent). The curves represent the population mean trajectory estimated using penalized fractional polynomials for each of three treatments. The light bands show 95% confidence intervals for the mean curve. The smooth non-linear pattern captures rapid early growth, steady middle-childhood growth, and adolescent acceleration.

Clinical Interpretation

The PFP approach provides several clinically relevant insights. First, the smooth mean curve characterises typical growth trajectories for this patient population, providing a reference for clinical monitoring. Second, the wide prediction intervals acknowledge substantial inter-individual differences, cautioning against rigid growth standards. Third, the REML-estimated smoothing parameter balances flexibility and parsimony, avoiding both over-smoothing (which would miss non-linear patterns) and under-smoothing (which would fit noise). Finally, confidence intervals enable hypothesis

testing about population growth rates at specific ages, while prediction intervals support individualised growth monitoring.

3.4.2 Application 2: Dermatology Clinical Trial

Study Design and Data

We analyze data from a randomized clinical trial comparing two treatments for scleroderma, a chronic autoimmune disease characterized by skin thickening. The primary outcome is change from baseline in modified Rodnan skin thickness score (MRSS), a validated measure of skin involvement. The dataset includes $N = 78$ patients (39 per treatment group) followed for up to 900 days, with 5–15 measurements per patient (median = 8). The scientific question is whether the treatments differ in their effects on skin thickness trajectories over time. Traditional analyses using linear mixed models with polynomial time effects may not adequately capture non-linear treatment responses, potentially missing clinically important differences.

Model Specification

We fit a PFP model with treatment-by-curve interactions:

$$\begin{aligned} y_{ij} = \beta_0 + \beta_1 \tilde{t}_{ij} + \gamma_0 \, I(\mathrm{trt}_i = 2) + \gamma_1 \, \tilde{t}_{ij} \, I(\mathrm{trt}_i = 2) \\ + \sum_{k=1}^{8} u_{g(i)k} \, \phi_{p_k}(\tilde{t}_{ij}) + v_i + w_i \tilde{t}_{ij} + \epsilon_{ij}, \end{aligned} \tag{3.23}$$

where y_{ij} is change from baseline in skin thickness for patient i at visit j, $\tilde{t}_{ij}$ is centred and scaled time (days since first visit), $I(\mathrm{trt}_i = 2)$ indicates treatment group 2 (reference = treatment 1), γ_0 and γ_1 represent treatment differences in intercept and linear slope, $u_{g(i)k}$ are treatment-specific random effects for FP basis k with $u_{gk} \sim N(0, \sigma_u^2)$ independently for $g = 1, 2$, and v_i, w_i are patient-specific random intercept and slope. This specification allows treatment-specific non-linear patterns

while sharing the smoothing parameter λ across treatments.

Results

Table 3.2 presents variance component estimates. The treatment-specific non-linear variance $\hat{\sigma}_u^2 = 0.0284$ is moderate relative to the residual variance $\hat{\sigma}_\epsilon^2 = 2.1609$, yielding $\hat{\lambda} = 76.1$. This indicates less smoothing than in the growth curve example, allowing more flexible treatment-specific patterns.

Table 3.2: Variance component estimates for the dermatology trial model.

Component	Estimate	Std. Error
Treatment-specific FP SD (σ_u)	0.1686	0.0312
Patient intercept SD (σ_v)	4.3250	0.4891
Patient slope SD (σ_w)	0.0875	0.0201
Intercept–slope correlation (ρ_{vw})	−0.412	0.089
Residual SD (σ_ϵ)	1.4703	0.0821
Smoothing parameter ($\hat{\lambda}$)	76.1	–

Fixed effects estimates (Table 3.3) reveal significant treatment differences. Treatment 2 shows a more favourable trajectory, with a larger negative linear slope ($\hat{\gamma}_1 = -0.82$, $p = 0.003$), indicating greater improvement over time.

Table 3.3: Fixed effects estimates for the dermatology trial model.

Parameter	Estimate	Std. Error	t-value	p-value
Intercept (β_0)	−0.23	0.52	−0.44	0.660
Time (β_1)	−0.35	0.11	−3.18	0.002
Treatment 2 (γ_0)	0.18	0.73	0.25	0.805
Treatment 2 × Time (γ_1)	−0.82	0.27	−3.04	0.003

Figure 3.2 displays treatment-specific mean curves with 95% confidence intervals. Both treatments show improvement (negative change from baseline) over time, but Treatment 2 demonstrates a steeper and more sustained decline. Importantly, the non-linear patterns differ between treatments: Treatment 1 shows early improvement

that plateaus after 400 days, while Treatment 2 exhibits continued improvement throughout the follow-up period.

To quantify treatment differences, we computed the proportion of the follow-up period where treatment confidence intervals do not overlap. Overall, 58% of time points show non-overlapping intervals. In the late follow-up period (days 675–900), 72% of time points show significant differences, indicating increasing treatment separation over time. At the study endpoint (day 900), Treatment 2 shows a mean change of -5.2 units (95% CI: -6.8 to -3.6) compared to -2.1 units for Treatment 1 (95% CI: -3.5 to -0.7). The difference of -3.1 units (95% CI: -5.3 to -0.9) is clinically meaningful, as a 3-unit reduction in MRSS is considered a minimal clinically important difference in scleroderma research (Khanna et al., 2019).

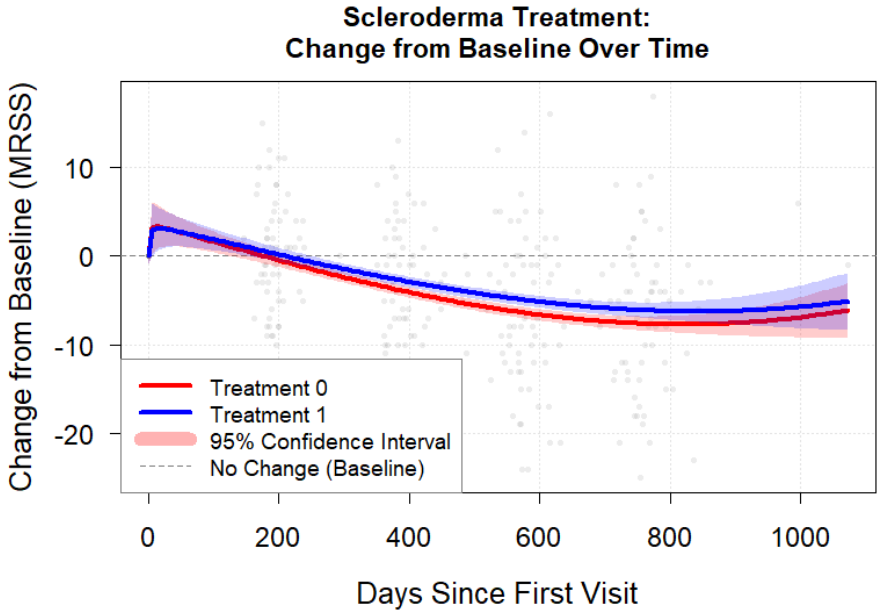

Figure 3.2: Treatment comparison in dermatology clinical trial. Change from baseline in skin thickness score over time for two treatment groups. Solid lines show treatment-specific mean curves estimated using penalized fractional polynomials with treatment-by-curve interactions. Shaded bands show 95% confidence intervals for mean curves. Points show individual patient measurements (semi-transparent). Treatment 0 (red) demonstrates greater improvement than Treatment 1 (blue), with increasing separation over time. The horizontal dashed line indicates no change from baseline.

Clinical Interpretation

The PFP analysis provides clinically actionable insights. Both treatments improve skin thickness, but Treatment 2 demonstrates superior efficacy, particularly in the long term. The non-linear patterns reveal that treatment differences emerge gradually and increase over time, suggesting that long-term follow-up is essential for capturing full treatment effects. The prediction intervals account for patient-level variability, enabling individualised treatment monitoring and identification of non-responders. Furthermore, the plateau in Treatment 1's effect after 400 days suggests a potential biological mechanism (e.g., tachyphylaxis or ceiling effect) warranting further investigation.

3.5 Discussion

3.5.1 Summary

We proposed penalized fractional polynomials implemented via linear mixed models for non-linear longitudinal data. The approach combines the interpretability of fractional polynomials with automatic smoothing via REML estimation of variance components. Applications to pediatric growth curves and a dermatology trial demonstrated the method's flexibility, interpretability, and integration with standard software.

3.5.2 Advantages

The proposed approach offers several advantages:

1. **Automatic Smoothing:** The smoothing parameter is estimated via REML rather than selected subjectively, providing an objective, data-driven approach.

2. **Interpretability:** Fractional polynomial parameterization is more interpretable than basis expansions like splines, particularly for monotonic or simple non-linear relationships.

3. **Software Integration:** Implementation uses standard mixed model software, making the method accessible to practitioners familiar with longitudinal data analysis.

4. **Flexibility:** The framework naturally accommodates treatment comparisons, repeated measures, and hierarchical structures common in biomedical research.

5. **Inference:** Standard mixed model inference provides confidence bands, hypothesis tests, and prediction intervals without additional computational burden.

3.5.3 Limitations and Extensions

Several limitations suggest directions for future work:

1. **Power Selection:** We specified powers a priori. Adaptive power selection within the penalized framework could further enhance flexibility.

2. **Multivariate Outcomes:** The current approach handles univariate responses. Extension to multivariate longitudinal data would broaden applicability.

3. **Non-Gaussian Outcomes:** We focused on continuous, normally distributed outcomes. Generalized linear mixed models could accommodate binary, count, or other outcome types.

4. **Computational Scalability:** For very large datasets, REML estimation may become computationally intensive. Approximate methods or sparse matrix techniques could improve scalability.

3.5.4 Practical Recommendations

Based on our experience, we offer the following guidance:

1. **Power Selection:** Start with a rich set of powers (e.g., $\{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$). The penalization will automatically downweight unnecessary powers.

2. **Basis Standardization:** Always standardize fractional polynomial bases to avoid numerical issues and improve REML convergence.

3. **Model Diagnostics:** Check residual plots, random effect distributions, and influential cases as in standard mixed model analysis.

4. **Software Choice:** The `nlme` package in R provides flexible specification of nested random effects. The `lme4` package offers faster computation for large datasets.

3.5.5 Conclusion

Penalized fractional polynomials via mixed models provide a practical, interpretable approach to non-linear longitudinal data analysis. By leveraging the mixed model framework, the method achieves automatic smoothing, accommodates repeated measures, and integrates with familiar software. Applications in pediatric growth and clinical trials demonstrate its utility for biomedical research.

Chapter 4

Comparison of Machine Learning Models for Hospital Readmission Prediction

4.1 Abstract

Background: Affecting over 25.5 million individuals in the United States, diabetes is linked to increased hospital readmission rates almost double that of patients without diabetes. Previous research suggested the predictive superiority of deep learning (DL) models, particularly long short-term memory (LSTM) models, compared to traditional machine learning (ML) models, in forecasting diabetes-related hospital readmissions. The objective of this study was to compare one DL and 10 ML methods for predicting 30-day readmission rates in patients with diabetes.

Methods: The dataset was obtained from "Diabetes 130-US Hospitals for Years 1999-2008" in the University of California Irvine (UCI) Machine Learning Repository and consisted of 101,766 unique encounters from 130 hospitals and integrated delivery networks across the United States. Encounters were included if they were an inpatient

encounter (i.e., hospital admission) during which diabetes was entered in the system as a diagnosis. Predictors of readmission included demographic variables (age, race, gender), admission and discharge type, number of procedures and medications, diagnosis, and hemoglobin A1c (HbA1c) testing, a measure of the last 3 months of glycemic control. The outcome was the rate of readmission within 30 days. This study built upon traditional ML algorithms by incorporating the Grey Wolf Optimizer (GWO), a swarm intelligence-based optimization algorithm, for feature selection. Data preprocessing involved handling missing values, encoding categorical variables, and converting the target variable into a binary classification task. Feature engineering included creating service utilization variables, age group midpoints, and grouping admission types. Categorical features were one-hot encoded, and the dataset underwent standard scaling. Feature selection was conducted using the GWO algorithm, and class imbalance was addressed with synthetic minority oversampling technique (SMOTE) during group k-fold cross-validation. Eleven ML algorithms, including random forest (RF), extreme gradient boosting (XGBoost), decision tree, support vector machine (SVM), and LSTM were employed for model training and evaluation.

Results: RF emerged as the most favorable model, consistently outperforming others in F1 score, accuracy, precision score, and recall score. The RF model achieved the highest F1 score (0.83) and accuracy (0.88), indicating its superior predictive ability, especially in balancing precision and recall. XGBoost showed comparable results, achieving the second highest F1 score (0.84) and accuracy (0.88).

Conclusions: The RF and XGBoost models with GWO outcompeted previous DL predictive modeling in diabetes hospital readmission scenarios. However, ML with GWO adoption should be carefully considered based on specific hospital needs, patient populations, and available resources, acknowledging ongoing advancements in predictive analytics for healthcare.

Keywords: Diabetes; hospital readmission; machine learning; Grey Wolf Opti-

mizer; feature selection

4.2 Introduction

Diabetes affects over 25.5 million individuals in the United States, representing a significant public health challenge (Parker et al., 2024). Patients with diabetes experience hospital readmission rates nearly double those of patients without diabetes, creating substantial clinical and economic burdens (Rubin, 2018; Gregory et al., 2018). Hospital readmissions within 30 days of discharge serve as a key quality metric and are associated with increased healthcare costs, patient morbidity, and mortality (Jiang et al., 2003).

Predicting which patients are at high risk for readmission enables targeted interventions, including enhanced discharge planning, medication reconciliation, patient education, and post-discharge follow-up (Gregory et al., 2018). Machine learning (ML) approaches offer promise for developing accurate prediction models that can identify high-risk patients and inform clinical decision-making.

4.2.1 Previous Work

Prior research has explored various ML and deep learning (DL) approaches for predicting diabetes-related hospital readmissions (Shang et al., 2021; Hai et al., 2023). Some studies suggested that DL models, particularly long short-term memory (LSTM) networks, outperform traditional ML methods due to their ability to capture complex temporal patterns and non-linear relationships in electronic health record (EHR) data (Hai et al., 2023).

However, DL models often require large datasets, substantial computational resources, and can lack interpretability - a critical consideration in clinical settings where understanding model predictions is essential for trust and adoption. Traditional

ML methods, particularly ensemble methods like random forest (RF) and gradient boosting, have demonstrated strong performance across many healthcare prediction tasks while maintaining greater interpretability (Shang et al., 2021; Thenappan et al., 2023).

4.2.2 Grey Wolf Optimizer

This study incorporates the Grey Wolf Optimizer (GWO), a swarm intelligence-based metaheuristic optimization algorithm inspired by the social hierarchy and hunting behavior of grey wolves (Mirjalili et al., 2014). GWO has shown effectiveness in feature selection tasks, helping identify the most relevant predictors while reducing dimensionality and computational burden (Kopecky and Dwyer, 2023).

By combining GWO-based feature selection with diverse ML algorithms, this study aims to achieve high prediction accuracy while maintaining model interpretability and computational efficiency (Cui et al., 2018).

4.2.3 Study Objectives

The primary objective of this study is to compare the performance of 11 ML algorithms—including one DL method (LSTM) and 10 traditional ML methods—for predicting 30-day hospital readmission in patients with diabetes (Strack et al., 2014). Specific aims include:

1. Implement GWO-based feature selection to identify the most predictive variables (Mirjalili et al., 2014)

2. Train and evaluate 11 ML algorithms using consistent preprocessing and validation procedures (Chawla et al., 2002)

3. Compare model performance using multiple metrics (F1 score, accuracy, precision, recall, AUC)

4. Identify the optimal algorithm(s) for clinical implementation

5. Provide interpretable insights into key readmission predictors (Moss et al., 1999)

4.3 Methods

4.3.1 Data Source

The dataset was obtained from the UCI Machine Learning Repository: "Diabetes 130-US Hospitals for Years 1999-2008" (Strack et al., 2014). This publicly available dataset contains 101,766 unique hospital encounters from 130 hospitals and integrated delivery networks across the United States. The data span 10 years (1999–2008) and represent inpatient encounters where diabetes was documented as a diagnosis.

4.3.2 Inclusion and Exclusion Criteria

Encounters were included if they met the following criteria:

- Inpatient hospital admission (not emergency department visit or outpatient encounter)
- Diabetes documented as a diagnosis (any position)
- Length of stay between 1 and 14 days
- Laboratory tests performed during encounter
- Medications administered during encounter

4.3.3 Predictor Variables

The dataset includes 50 predictor variables across several categories:

Demographics:

- Age (grouped into 10-year intervals)
- Gender
- Race/ethnicity

Admission Characteristics:

- Admission type (emergency, urgent, elective, etc.)
- Admission source (physician referral, emergency department, transfer, etc.)
- Time in hospital (days)

Clinical Variables:

- Primary diagnosis (ICD-9 code)
- Secondary diagnoses (ICD-9 codes)
- Number of diagnoses
- Medical specialty of admitting physician

Procedures and Tests:

- Number of lab procedures
- Number of procedures (other than lab)
- Number of medications
- Number of outpatient visits in year prior to encounter
- Number of emergency visits in year prior to encounter
- Number of inpatient visits in year prior to encounter

Diabetes-Specific Variables:

- HbA1c test result (normal, abnormal, not measured)
- Diabetes medication prescribed (binary indicator across 24 generic medications)
- Medication dosage changes during encounter (up, down, steady, or not prescribed)

Discharge Information:

- Discharge disposition (home, skilled nursing facility, etc.)

4.3.4 Outcome Variable

The primary outcome is hospital readmission within 30 days of discharge. This was coded as a binary variable:

- 1 = Readmitted within 30 days
- 0 = Not readmitted within 30 days (including readmission after 30 days and no readmission)

4.3.5 Data Preprocessing

Missing Data and Initial Processing

Several preprocessing steps were applied to prepare the dataset:

- Columns with high proportions of missing values or only one unique value were removed
- Duplicate patient records were eliminated based on the `patient_nbr` column, retaining 101,766 unique encounters
- Admission type, discharge type, and admission source were re-encoded into fewer categories

Feature Engineering

Several derived features were created:

- **Service utilization score:** Sum of prior outpatient, emergency, and inpatient visits
- **Age midpoint:** Numeric representation of age intervals (e.g., [50–60) = 55)
- **Diagnosis grouping:** Primary diagnoses mapped to clinically meaningful numerical categories (circulatory, respiratory, digestive, etc.)
- **Total medications:** Calculated as the total number of medications used by each patient

Following feature engineering, the dataset contained 55 features across 101,776 encounters.

Encoding Categorical Variables

Categorical variables were one-hot encoded, creating binary indicator variables for each category.

Feature Scaling

All numeric features were standardized using z-score normalization:

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma} \tag{4.1}$$

where μ is the mean and σ is the standard deviation of the feature.

4.3.6 Feature Selection: Grey Wolf Optimizer

The Grey Wolf Optimizer (GWO) is a metaheuristic optimization algorithm inspired by the leadership hierarchy and hunting mechanism of grey wolves in nature (Mirjalili

et al., 2014). The algorithm simulates the social hierarchy (alpha, beta, delta, and omega wolves) and collaborative hunting behavior to search for optimal solutions. GWO was implemented using the `zoofs` library (version 0.1.26) (Strack et al., 2014).

GWO Algorithm

The GWO algorithm operates as follows:

Initialization:

- Initialize a population of grey wolves (candidate feature subsets) randomly
- Each wolf represents a binary vector where 1 indicates feature inclusion and 0 indicates exclusion

Fitness Evaluation:

- Train a base classifier with the feature subset represented by each wolf
- Identify alpha (best), beta (second best), and delta (third best) solutions

Position Update:

- Update positions of omega wolves based on alpha, beta, and delta positions
- Apply hunting behaviour equations to balance exploration and exploitation

Output:

- Return the feature subset corresponding to the alpha wolf

Selected Features

GWO reduced the dataset from 55 features to 43 features. Key selected features included:

- Discharge disposition

- Service utilization score
- Number of inpatient visits (prior year)
- Time in hospital
- Age
- Primary diagnosis category
- Number of procedures
- Number of diagnoses
- Number of medications
- Number of lab procedures
- HbA1c test result
- Insulin and metformin use

4.3.7 Handling Class Imbalance

The dataset exhibits class imbalance, with 11,357 of 101,766 encounters (11.2%) resulting in 30-day readmission. To address this, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training set within each cross-validation fold to prevent data leakage (Chawla et al., 2002). SMOTE generates synthetic examples of the minority class by interpolating between existing minority class samples and their nearest neighbours, resulting in a balanced representation of readmitted and non-readmitted cases.

4.3.8 Machine Learning Algorithms

We evaluated 11 ML algorithms:

1. **Random Forest (RF):** Ensemble of decision trees; robust to overfitting and provides feature importance estimates

2. **Extreme Gradient Boosting (XGBoost):** Gradient boosting with regularization; excels on structured data and handles missing values

3. **Decision Tree:** Single tree with recursive binary splitting; interpretable but prone to overfitting

4. **Support Vector Machine — linear kernel (SVM):** Maximum margin classifier; effective in high-dimensional spaces

5. **Support Vector Machine — RBF kernel (SVM):** Handles non-linear decision boundaries; requires careful hyperparameter tuning

6. **K-Nearest Neighbors (KNN):** Instance-based learning; simple but computationally expensive at prediction time

7. **Naive Bayes:** Probabilistic classifier assuming feature independence; simple and efficient

8. **Logistic Regression:** Linear model with logistic link function; interpretable but limited for non-linear relationships

9. **AdaBoost:** Adaptive boosting combining weak learners; resilient against overfitting but sensitive to noise

10. **Multilayer Perceptron (MLP):** Feedforward neural network; captures complex non-linear relationships but requires careful tuning

11. **Long Short-Term Memory (LSTM):** Recurrent neural network with memory cells; suited to sequential data

4.3.9 Model Training and Validation

Cross-Validation

We used group k-fold cross-validation with 5 folds, grouping all encounters for the same patient within the same fold to avoid data leakage across patient visits. SMOTE was applied exclusively to the training data within each fold. Performance metrics were averaged across the 5 folds to obtain stable, less biased estimates of model performance.

Hyperparameter Settings

Default hyperparameters were used for most traditional ML models. The following exceptions applied:

- **MLP and Logistic Regression:** Maximum of 1,000 iterations specified
- **XGBoost:** Configured to use log loss as the evaluation metric
- **LSTM:** A sequential model with two LSTM layers of 50 units each, followed by dropout layers (rate = 0.2), a dense layer with sigmoid activation, binary cross-entropy loss, and the Adam optimizer; trained for 10 epochs with a batch size of 32

4.3.10 Performance Metrics

Models were evaluated using multiple metrics:

- **F1 Score** (primary outcome): Harmonic mean of precision and recall

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.2}$$

- **Accuracy:** Proportion of correct predictions

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.3}$$

- **Precision:** Proportion of positive predictions that are correct

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.4}$$

- **Recall (Sensitivity):** Proportion of actual positives correctly identified

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.5}$$

- **AUC-ROC:** Area under the receiver operating characteristic curve

where TP = true positives, TN = true negatives, FP = false positives, FN = false negatives. 95% confidence intervals were computed across the 5 cross-validation folds for each metric.

4.3.11 Implementation

All analyses were conducted in Python using scikit-learn for traditional ML algorithms, XGBoost for gradient boosting, TensorFlow/Keras for neural networks (MLP and LSTM), and the `zoofs` library (version 0.1.26) for GWO-based feature selection (Mirjalili et al., 2014).

4.4 Results

4.4.1 Dataset Characteristics

Table 4.1 summarizes the characteristics of the 101,766 encounters included in the analysis. Of these, 11,357 (11.2%) resulted in readmission within 30 days. The majority of patients were over 60 years of age (over 65%), Caucasian (74.8%), and female (53.8%). Most patients were admitted through the emergency department (53.1%) and spent a mean of 4.4 days in hospital before being discharged home (59.2%). Only 16.7% of patients had an HbA1c test performed during their encounter.

4.4.2 Feature Selection Results

Following preprocessing and feature engineering, the dataset contained 55 features. GWO reduced this to 43 features. The most important features identified by the RF model using Gini gain were, in descending order: discharge disposition, service utilization score, number of inpatient visits, time in hospital, age, primary diagnosis category, number of procedures, number of diagnoses, number of medications, number of lab procedures, HbA1c result, insulin use, and metformin use (Figure 4.1).

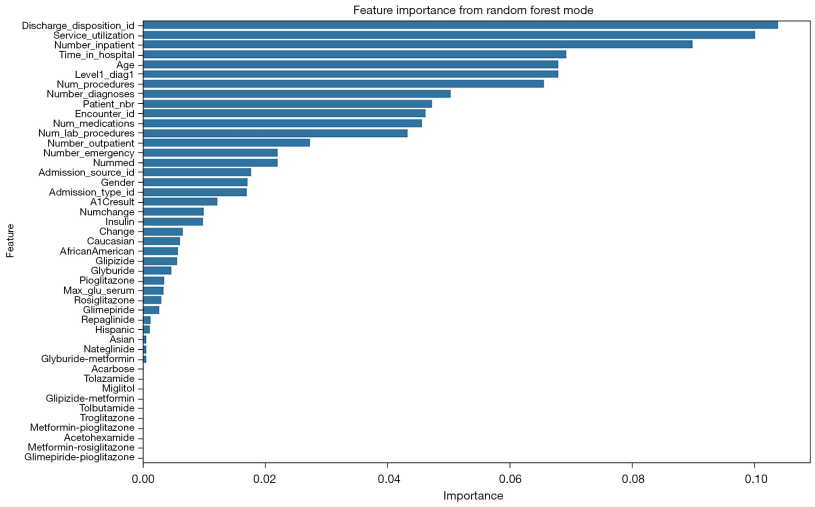

Figure 4.1: Feature importance for random forest analysis.

4.4.3 Model Performance Comparison

Table 4.2 presents the performance of all 11 ML algorithms, averaged across 5 cross-validation folds. The narrow 95% confidence intervals across all models underscore the reliability of the estimates.

4.4.4 Key Findings

Top Performers

RF, XGBoost, and AdaBoost emerged as the top-performing algorithms.

- **Random Forest:** Consistently outperformed other models across F1 score (0.83), accuracy (0.88), precision (0.83), and recall (0.88). RF demonstrated superior ability to distinguish between positive and negative classes, with an AUROC of 0.63, and showed a robust balance between precision and recall.
- **XGBoost:** Achieved the highest F1 score (0.84) and matched RF on accuracy (0.88), with a marginally higher AUROC (0.64). XGBoost showed comparable performance to RF across all metrics and represents a strong competitive alternative.
- **AdaBoost:** Also achieved an F1 score of 0.84, equal to XGBoost, with accuracy of 0.86 and AUROC of 0.62.

Deep Learning Performance

Contrary to some previous reports (Hai et al., 2023), the LSTM model did not outperform traditional ML methods. LSTM achieved an F1 score of 0.79 and accuracy of 0.77, comparable to MLP but below RF, XGBoost, and AdaBoost. The LSTM also exhibited wider confidence intervals across folds (F1: 0.77–0.81; accuracy: 0.74–0.80), suggesting greater instability relative to the ensemble methods.

This may be attributed to:

1. Limited temporal structure in the data (single hospitalization encounter rather than a longitudinal time series)

2. Smaller dataset size relative to Hai et al. (2023), who used over 2.8 million encounters, which may have allowed their DL model to perform better

3. Lack of sequential dependencies that LSTM architectures are specifically designed to exploit

4.4.5 Feature Importance Analysis

Figure 4.1 shows features ranked by importance in the RF model using Gini gain. Discharge disposition was the single most important predictor, followed by service utilization score and number of inpatient visits. Demographic factors were also important, in the order of age, gender, and ethnicity. Details of the initial hospitalisation were similarly predictive, including time in hospital, number of procedures, number of medications, and number of lab procedures. Among diabetes-specific variables, HbA1c result, insulin use, and metformin use were important in that order. Specific medications ranging from glipizide to metformin-rosiglitazone were of lesser predictive importance.

The partial dependence plot (Figure 4.2) shows that higher levels of both HbA1c and serum glucose were associated with a higher likelihood of readmission.

4.4.6 ROC Curves

ROC curves were plotted for RF, XGBoost, and LSTM (Figure 4.3). RF and XGBoost achieved AUROCs of 0.63 (95% CI: 0.63, 0.64) and 0.64 (95% CI: 0.64, 0.65) respectively, both superior to the LSTM model (0.61, 95% CI: 0.61, 0.61). The proximity of

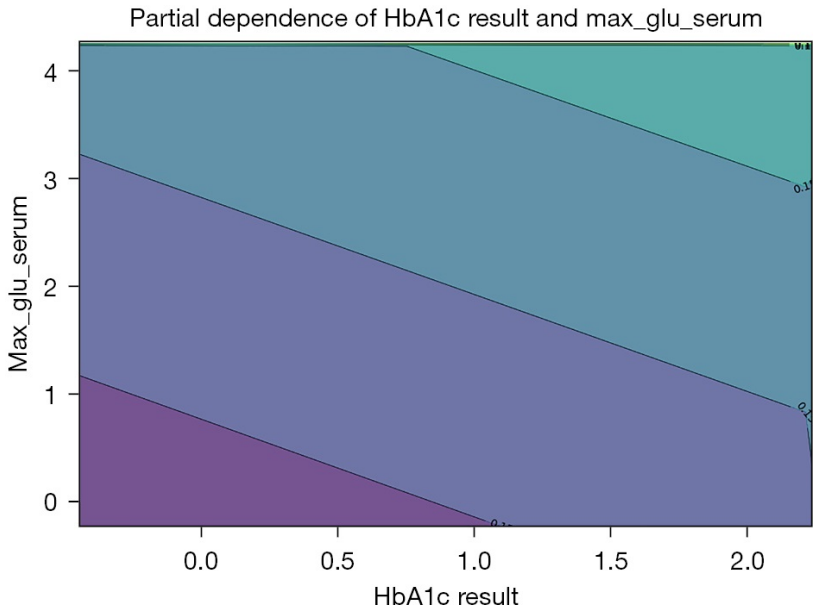

Figure 4.2: Partial dependence plots for two variables from the random forest model.

the RF and XGBoost ROC curves to the upper left corner indicates their stronger ability to correctly distinguish positive from negative cases relative to LSTM.

4.5 Discussion

4.5.1 Summary of Findings

This study compared 11 ML algorithms for predicting 30-day hospital readmission in patients with diabetes. RF and XGBoost, combined with GWO-based feature selection, emerged as the top-performing methods, achieving F1 scores of 0.83–0.84 and accuracy of 0.88. These results demonstrate that traditional ML methods, when enhanced with intelligent feature selection and class imbalance handling, can match or outperform deep learning approaches for this prediction task (Shang et al., 2021; Hai et al., 2023).

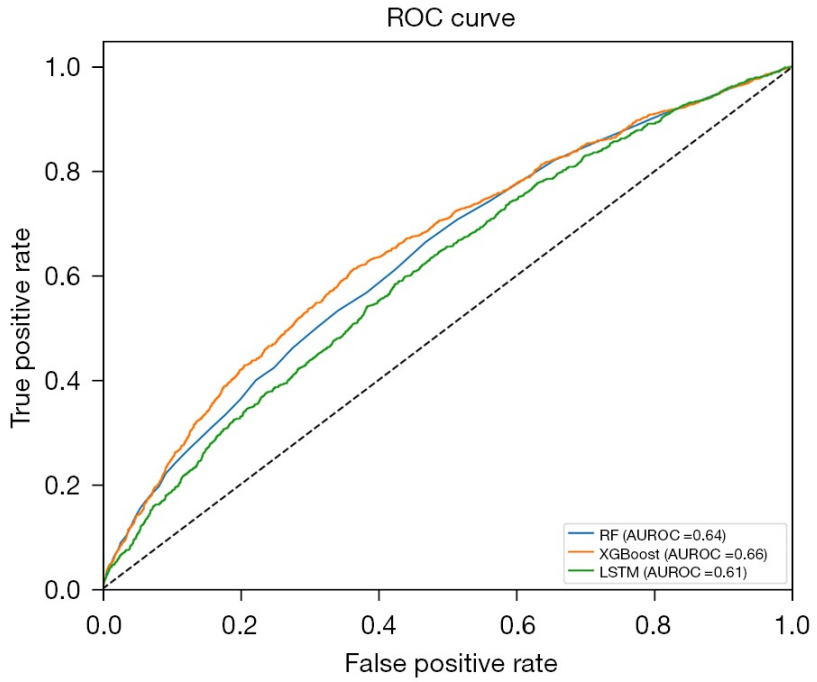

Figure 4.3: ROC curve for readmission prediction for three trained models.

4.5.2 Comparison to Previous Literature

Our results contrast with some prior studies that reported superior performance of deep learning models, particularly LSTM networks. Hai et al. (2023) found that LSTM achieved an AUROC of 0.79 compared to RF at 0.72 ($P<0.0001$). However, our GWO-optimised RF and XGBoost outperformed the LSTM model in this study. Several factors may explain these discrepancies:

1. **Data Structure:** Our dataset consists of single-encounter data with limited temporal sequencing. LSTM networks excel with true time series data featuring multiple encounters over time. The lack of sequential dependencies may have limited LSTM's advantage.

2. **Sample Size:** (Hai et al., 2023) used a dataset from the Temple University Health System encompassing over 2,836,569 encounters, which likely allowed the DL model to perform better. While 101,766 encounters is substantial for traditional ML, it may be insufficient for deep learning to reach optimal

performance.

3. **Feature Engineering:** Our GWO-based feature selection provided traditional ML methods with a refined, high-quality feature set, which may have reduced the advantage of deep learning's automatic feature learning.

Our findings of RF superiority are consistent with Shang et al. (2021), who also found RF to be the best-performing classifier across over 100,000 diabetes records, and with Thenappan et al. (2023), whose RF model achieved 83% precision. Our algorithms also outperformed Rajput and Alashetty (2022), who achieved 77.4% accuracy, and Cui et al. (2018), who achieved 81.02% accuracy with an improved SVM method. Our findings of XGBoost as a competitive alternative are consistent with Cuong and Wang (Wang et al., 2021), who found XGBoost to be the best-performing model in a similar comparison.

4.5.3 Clinical Implications

Actionable Predictors

Several identified predictors are potentially modifiable through clinical interventions:

- **Diabetes medication changes:** Ensuring patients understand medication changes made during hospitalisation and have appropriate follow-up may reduce readmissions.
- **HbA1c testing:** The association between HbA1c testing and lower readmission risk suggests that inpatient HbA1c measurement should be routine for diabetes patients. Strack et al. (2014) found that HbA1c monitoring appeared to be associated with lower readmission rates, particularly when diabetes was the primary diagnosis.

- **Discharge disposition:** Discharge type was the single most important predictor of readmission in the RF model, underscoring the importance of appropriate post-acute care placement.

- **Service utilisation:** Prior inpatient and emergency visits were among the strongest predictors, emphasising the importance of longitudinal data and continuity of care (Hai et al., 2023).

Health System Benefits

Higher hospital readmission rates are associated with increased disease burden for patients with diabetes and carry important medical, social, quality of life, and financial implications (Jiang et al., 2003). A modest 5% reduction in 30-day readmission rates has been estimated to result in 82,754 fewer admissions per year and annual cost savings of $1.2 billion (Rubin, 2018; Shang et al., 2021). Predictive models enabling targeted transitional care interventions - such as close outpatient follow-up, home health visits, and telemedicine follow-up shortly after discharge - could help achieve such reductions (Gregory et al., 2018).

4.5.4 Model Interpretability

A key advantage of RF and XGBoost over deep learning is interpretability. Feature importance analysis provides clinically meaningful insights:

- Prior healthcare utilisation (inpatient and emergency visits) emerged as among the strongest predictors, consistent with prior literature emphasising the importance of longitudinal data (Hai et al., 2023).

- Diagnostic complexity (number of diagnoses) reflects multimorbidity, a well-established readmission risk factor.

- Diabetes-specific factors (HbA1c result, insulin and metformin use) highlight opportunities for improved inpatient diabetes management.

4.5.5 Implementation Considerations

Integration into Clinical Workflow

For the study to be useful clinically, these models must be integrated into real-time hospital settings to predict patients most likely to be readmitted (Strack et al., 2014). Hospitals should adjust the positive class threshold for readmissions based on their specific needs, balancing more false positives against fewer false negatives to find the optimal threshold.

Ethical Considerations

Deployment of prediction models raises important ethical concerns, including the potential for algorithmic bias across demographic subgroups. Careful validation in diverse populations is essential, particularly given the limited Hispanic representation and the restricted length-of-stay range (1–14 days) in the current dataset.

4.5.6 Limitations

Data Limitations

- **Age of Data:** The dataset spans 1999–2008, and healthcare practices, treatment guidelines, and diabetes medications have evolved substantially. Model performance in contemporary settings requires validation.
- **Single Data Source:** Although the dataset covers 130 hospitals across four US geographic regions (Midwest, Northeast, South, and West), it may not generalise to other healthcare systems or countries. The sample may not be representative of more recent patients with diabetes.

- **Missing Variables:** Important predictors such as social determinants of health are not captured. Additionally, predictors such as blood serum glucose can vary from day to day, limiting their reliability as static features.

- **Population Representativeness:** Hispanic representation in the dataset is low, and the length of stay was restricted to 1–14 days, which may limit generalisability to patients with very short or very prolonged admissions.

Methodological Limitations

- **Binary Outcome:** Readmission was dichotomised as within or after 30 days. Readmissions over a longer period are primarily related to chronic disease progression and reflect outpatient care quality, whereas early readmissions reflect quality of the initial admission (McKay et al., 1997; Benbassat and Taragin, 2000).

- **Single Timepoint:** The model uses data from a single hospitalisation encounter. Longitudinal models incorporating multiple encounters have shown improved performance in other work (Hai et al., 2023).

- **Causality:** Observational data cannot establish causal relationships. Identified predictors may be markers rather than causes of readmission risk.

4.5.7 Future Directions

- Apply the trained models to patients with diabetes in a real-time hospital setting to identify those most at risk for readmission based on their demographics, admission status, and laboratory results.

- Incorporate additional clinical parameters such as hypoglycaemia and hyperglycaemia events, point-of-care glucose testing, and continuous glucose monitoring,

both in patients with and without diabetes (Zapatero et al., 2014; Cichosz and Schaarup, 2017).

- Validate models in more contemporary datasets and across diverse healthcare systems and geographic regions.
- Investigate the effects of including longitudinal encounter data, which have been shown to improve DL model performance (Hai et al., 2023).

4.5.8 Conclusion

This study demonstrates that RF and XGBoost algorithms, combined with GWO-based feature selection, group k-fold cross-validation, SMOTE, and feature engineering, produce models with F1 scores rivalling or outperforming those in the existing literature. These traditional ML methods outperformed the LSTM model evaluated here, demonstrating that deep learning is not the only viable approach for predicting diabetes-related hospital readmissions.

However, adoption of ML with GWO should be carefully considered based on specific hospital needs, patient populations, and available resources. Other advanced algorithms, including deep learning, still play a pivotal role in predictive analytics for healthcare (Hai et al., 2023). Hospitals are encouraged to begin using these models to improve patient care and ultimately reduce diabetes-related comorbidities and complications, while recognising the ongoing advancements in the field.

Table 4.1: Dataset characteristics for diabetes readmission study.

Characteristic	Overall, n (%)	Readmitted <30 days, n (%)
Total encounters	101,766 (100.0)	11,357 (100.0)
Gender		
Female	54,708 (53.8)	6,152 (54.2)
Male	47,055 (46.2)	5,205 (45.8)
Race		
Caucasian	76,099 (74.8)	8,592 (75.7)
African American	19,210 (18.9)	2,155 (19.0)
Hispanic	2,037 (2.0)	212 (1.9)
Asian	641 (0.6)	65 (0.6)
Other	1,506 (1.5)	145 (1.3)
Admission Type		
Emergency	53,990 (53.1)	6,221 (54.8)
Urgent	18,480 (18.2)	2,066 (18.2)
Elective	18,869 (18.5)	1,961 (17.3)
Discharge Disposition		
Home	60,234 (59.2)	5,602 (49.3)
Skilled nursing facility	13,954 (13.7)	2,046 (18.0)
Home health service	12,902 (12.7)	1,638 (14.4)
Clinical Characteristics (mean)		
Time in hospital (days)	4.4	4.8
Number of diagnoses	7.4	7.7
Number of procedures	1.3	1.3
Number of medications	16.0	16.9
Number of lab procedures	43.1	44.2
HbA1c Result		
>8%	8,216 (8.1)	811 (7.1)
>7%	3,812 (3.7)	383 (3.4)
Normal (4.0–5.6%)	4,990 (4.9)	482 (4.2)
Not tested	84,748 (83.3)	9,681 (85.2)
Diabetes Medications		
Prescribed	78,363 (77.0)	9,111 (80.2)
Not prescribed	23,403 (23.0)	2,246 (19.8)

Table 4.2: Performance comparison of 11 machine learning algorithms for 30-day readmission prediction. Results shown as mean (95% CI) across 5-fold cross-validation.

Algorithm	F1 Score	Accuracy	Precision	Recall	AUC-ROC
XGBoost	0.84 (0.83, 0.84)	0.88 (0.88, 0.89)	0.83 (0.83, 0.84)	0.88 (0.88, 0.89)	0.64 (0.64, 0.65)
AdaBoost	0.84 (0.83, 0.84)	0.86 (0.85, 0.86)	0.82 (0.82, 0.82)	0.86 (0.85, 0.86)	0.62 (0.61, 0.63)
Random Forest	0.83 (0.83, 0.84)	0.88 (0.88, 0.89)	0.83 (0.82, 0.83)	0.88 (0.88, 0.89)	0.63 (0.63, 0.64)
Decision Tree	0.80 (0.80, 0.80)	0.79 (0.79, 0.79)	0.81 (0.80, 0.81)	0.79 (0.79, 0.79)	0.52 (0.52, 0.53)
MLP	0.79 (0.78, 0.80)	0.77 (0.77, 0.78)	0.81 (0.81, 0.82)	0.77 (0.77, 0.78)	0.58 (0.57, 0.59)
LSTM	0.79 (0.77, 0.81)	0.77 (0.74, 0.80)	0.82 (0.82, 0.82)	0.77 (0.74, 0.80)	0.61 (0.61, 0.61)
SVM RBF kernel	0.74 (0.74, 0.74)	0.69 (0.69, 0.69)	0.82 (0.82, 0.82)	0.69 (0.69, 0.69)	0.48 (0.46, 0.49)
Logistic Regression	0.70 (0.70, 0.70)	0.64 (0.63, 0.64)	0.83 (0.83, 0.83)	0.64 (0.63, 0.64)	0.63 (0.62, 0.63)
SVM linear kernel	0.70 (0.70, 0.70)	0.63 (0.63, 0.63)	0.83 (0.83, 0.83)	0.63 (0.63, 0.63)	0.49 (0.46, 0.52)
KNN	0.69 (0.69, 0.69)	0.63 (0.62, 0.63)	0.81 (0.81, 0.81)	0.63 (0.62, 0.63)	0.55 (0.55, 0.55)
Naïve Bayes	0.03 (0.03, 0.03)	0.12 (0.12, 0.12)	0.82 (0.77, 0.86)	0.12 (0.12, 0.12)	0.50 (0.50, 0.50)

Chapter 5

Optimizing Blood Glucose Predictions Using Advanced Forecasting Techniques

5.1 Introduction

Diabetes is a leading cause of morbidity and mortality worldwide. The most recent World Health Organization (WHO) data estimate a prevalence of 828 million adults aged 18 and older in 2022, representing an increase of 630 million since 1990 (Zhou et al., 2024). Type 1 diabetes (T1D) accounts for approximately 2% of all diabetes, with variations in prevalence from less than 1% in China to 17% in Finland (Green et al., 2021).

Public health surveillance is essential for understanding the worldwide burden of diabetes (Liu et al., 2024). Big data collected from patients has become increasingly available, enabled by wearable medical devices and other technology, offering new opportunities for healthcare professionals to diagnose, treat, and manage diabetes and its complications, as well as for patients in their diabetes self-management (Fagherazzi

and Ravaud, 2019; Riddle et al., 2019; Kavakiotis et al., 2017). Optimizing blood glucose values is critical for the prevention of diabetes-associated complications and mortality (Stratton et al., 2000).

The typical target blood glucose range for continuous glucose monitoring (CGM) is 70–180 mg/dL, with a goal to spend at least 70% of the time within this range (Committee et al., 2023). Controlling blood glucose levels and achieving this time in range can be challenging due to multiple factors, including diet, exercise, medication side effects, variable patient literacy, financial constraints, nonadherence, stress, and illness (Chepulis et al., 2021).

5.1.1 Continuous Glucose Monitoring and Predictive Algorithms

CGM devices consist of a sensor placed in the subcutaneous tissue of the arm, abdomen, or lower back, and a transmitter. They are classified as intermittently scanned CGM or real-time CGM (Shaw et al., 2024) and measure glucose levels in the interstitial fluid every 1–15 minutes for up to 15 days (Zhu et al., 2020). These CGM data provide a view of glucose trends over time, and CGM has become the standard of care in T1D (Committee et al., 2023).

Integrating predictive algorithms into CGM devices or cellular applications allows for real-time glucose predictions (Ying et al., 2025). The artificial pancreas and insulin bolus calculators are examples of systems that use machine learning based on CGM data to assist with improving glycemic control (Zhu et al., 2020). These predictions can use state-of-the-art algorithms including time series analyses, physiological models, machine learning, and deep learning.

5.1.2 Time Series Approaches

Several time series techniques are available for blood glucose prediction. Auto-Regressive Integrated Moving Average (ARIMA) models require the selection of three parameters: p (autoregressive order), d (differencing order), and q (moving average order). Algorithms have been developed that adaptively determine these orders and estimate parameters based on previous blood glucose data (Yang et al., 2018). These adaptive ARIMA models, along with Bayesian approaches, represent the current state of the art in blood glucose prediction (QingXiang et al., 2023). However, applying other types of time series analysis - particularly Long Short-Term Memory (LSTM) networks and other machine learning algorithms - to blood glucose prediction may prove more efficient under the right conditions (Goel and Sharma, 2022).

5.1.3 Physiological Models

The original physiological models employed to predict blood glucose used a greater degree of human input than typically used in time series and machine learning-generated prediction. These models were based on differential equations modelling two key processes: insulin kinetics and the effects of insulin and glucose after intravenous glucose injection (Bergman, 2021). The key parameters determined to be of greatest importance in these early studies were insulin, glucose, the liver, and peripheral tissues.

In contrast to the parsimony of "minimal" models, "maximal" physiological models seek to manually incorporate as many relevant physiological processes as possible (Pompa et al., 2021). One such model is the Hovorka physiological model, an adaptive nonlinear model that uses Bayesian techniques to estimate time-varying parameters for a complex compartment model of the glucoregulatory system (Hovorka et al., 2004). Machine learning can be used to optimise the parameters of both time series and physiological models of blood glucose in T1D patients, creating hybrid models (Yang et al., 2018; Hovorka et al., 2004).

5.1.4 Machine Learning Approaches

The list of machine learning algorithms available for blood glucose prediction is vast; however, some of the most common include decision trees, Random Forest, support vector machines, and k-nearest neighbours. LSTM, XGBoost, Support Vector Regression, and Multilayer Perceptron are among the most accurate (Gómez-Castillo et al., 2021; Fu et al., 2023; Mayo et al., 2019). Similarly to older non-automated models, the parameters of machine learning models can be optimised and, interestingly, non-automated time series algorithms can preprocess the data to improve the accuracy of these models (Gómez-Castillo et al., 2021).

5.1.5 Study Objectives

This study opted to use the Grey Wolf Optimizer (GWO) for optimisation of machine learning parameters (Mirjalili et al., 2014), rather than expert-driven optimisation. The primary objective was to compare commonly used machine learning models along with their stacked counterpart. Stacking combines two or more learners into a meta-model that leverages each individual model's prediction to produce an overall improved prediction (Lamari et al., 2024). Stacking often increases predictive accuracy, though not always - for example, when the base learners are poor models. The study further aimed to determine the importance of including bolus measurements to aid in time series prediction, since it can be costly to collect such data and their inclusion may not always improve predictive accuracy.

The study was also novel in its use of Grey Wolf Optimization for hyperparameter tuning. Grid search and random search are two common tuning methods: the first tests every possibility, which can be time-consuming, while the second is not guaranteed to converge on an optimal solution. GWO remedies these deficiencies by converging on an optimum (Mirjalili et al., 2014). To our knowledge, GWO had not previously been applied to CGM prediction.

The study was motivated by the AWD Stacking Ensemble learner of Yang et al. (2024), which demonstrated that an advanced stacking ensemble method outperformed simpler learners including StackLSTM. While adaptive ARIMA has also shown promise in outperforming traditional non-optimised time series algorithms (Yang et al., 2018), ARIMA does not allow for the inclusion of bolus measurements - a gap this study aimed to address. Thus, this study sought to contribute to predictive analytics in personalised medicine by allowing for patient-specific features.

5.2 Methods

5.2.1 Dataset

The DiaTrend dataset (Prioleau et al., 2023) was used in this study. It consists of longitudinal data collected from wearable medical devices, including over 27,561 days of CGM and 8,220 days of insulin pump data from 54 patients with T1D. The CGM sheet contains two columns: date and mg/dL. Each subject also had a bolus sheet containing date, normal dose, carbohydrate input, insulin-to-carbohydrate ratio, blood glucose input, recommended carbohydrate, and recommended net dose. Only 17 of the 54 subjects had a bolus sheet. The DiaTrend dataset is useful for several research questions including blood glucose prediction, prediction of hypoglycemia and hyperglycemia, detection of unannounced meals, algorithm development for insulin delivery, and understanding adherence to wearable medical devices (Prioleau et al., 2023).

5.2.2 Demographic Evaluation

A binomial test was used to determine whether the observed proportion of females (68.5%) differed significantly from an expected proportion of 50%. The test was two-tailed, with the null hypothesis that the observed proportion equalled the expected

proportion.

For hemoglobin A1C, a two-tailed t-test was used to determine the significance of the difference between the observed mean and the hypothesised population mean ($\mu_0 = 5.7$), using 50 degrees of freedom from 51 of the 54 patients with available data. The hypothesised population mean was selected from the Mayo Clinic's healthy reference value of $< 5.7\%$. The p-value was calculated using the t-distribution.

5.2.3 Grey Wolf Parameter Optimization

Hyperparameter tuning was performed using a GWO algorithm implemented in Python 3.10.12 (Mirjalili et al., 2014). GWO mimics the social hierarchy and hunting behaviour of grey wolves to identify optimal solutions within a defined search space. It is a heuristic algorithm that can be more efficient than random search or grid search. Being heuristic, its convergence to a global minimum is not guaranteed, but it often performs well in complex optimisation problems including feature selection (Mirjalili et al., 2014).

Three discrete hyperparameter values were preselected for each machine learning model. A continuous mapping was defined to map the parameters of the GWO algorithm to each discrete value. The algorithm was run for 3 iterations on a GPU using 10 wolves. The pseudocode for the GWO algorithm is as follows:

1. Initialise wolf population within $[\min, \max]^d$ bounds
2. Evaluate fitness for all wolves using the target function
3. Identify alpha, beta, delta (top 3 fittest wolves)
4. For $t = 1$ to iterations:
   (a) $a \leftarrow 2 \times (1 - t/\text{iterations})$ (linear decay)

(b) For each wolf, update position: $X'(t) = (X_\alpha + X_\beta + X_\delta)/3 + a \cdot (\text{rand}() - 0.5)$ (collective hunting)

(c) Apply boundary constraints $[\min, \max]$

5. Update hierarchy (α, β, δ) and return best solution

5.2.4 Data Preprocessing

CGM and bolus data from 54 subject files were processed to generate a dataset suitable for machine learning predictive modelling. To incorporate temporal dependencies, lag features were generated for the CGM mg/dL column, spanning one to twelve time steps prior. The target variable was set to be the glucose measurement 15 minutes ahead. Since measurements were taken mostly every 5 minutes, this corresponded to a 3-step ahead prediction. A 15-minute prediction horizon was chosen as a reasonable window for notifying a user of an impending abnormality in glycemic management.

The bolus sheet was integrated by adding up to three bolus injections that occurred most recently before the target prediction time. Along with the bolus amount, features for insulin on board, insulin sensitivity, carbohydrate intake, and time elapsed since bolus administration were created. The dataset was split into training and testing subsets using the `TimeSeriesSplit` function in Python. To prevent data leakage across folds, each patient's data was split individually, and metrics were obtained using 5-fold cross-validation (Figure 5.1).

5.2.5 Machine Learning Algorithms

Three baseline machine learning algorithms were evaluated, along with one stacking ensemble meta-learner:

1. **Random Forest (RF):** Excels at handling lagged variables and capturing short-term patterns in time series data.

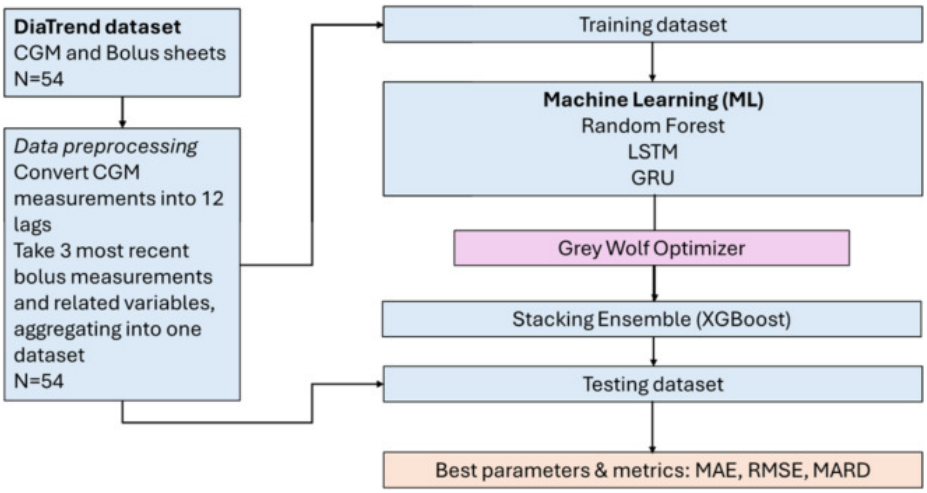

Figure 5.1: Study workflow for analysis.

2. **Long Short-Term Memory (LSTM):** A recurrent neural network architecture well-suited for modelling long-term dependencies in sequential data.

3. **Gated Recurrent Unit (GRU):** A recurrent neural network variant similar to LSTM but with a simplified gating mechanism; also well-suited for long-term temporal dependencies.

4. **XGBoost (stacking meta-learner):** Used to combine the predictions of the three baseline models into a final stacked prediction, selected for its balance of speed and performance in capturing nonlinear relationships.

5.2.6 Stacking Ensemble Framework

A patient-centric time-series validation framework was developed to ensure temporal integrity and clinical relevance. A custom `SafePatientTimeSeriesSplit` cross-validator was implemented, enforcing three key constraints: (1) strict temporal splitting where training data always precedes validation data within each patient's timeline; (2) minimum sample thresholds (20 training and 10 validation samples per patient) to ensure meaningful learning; and (3) patient-wise grouping to prevent data leakage between folds. The validation strategy splits patients into 5 chronological folds while

maintaining their internal temporal sequences, with validation sets containing only the most recent measurements from held-out patients.

XGBoost was employed with temporal early stopping (20-round patience on validation MAE) to prevent overfitting. Hyperparameters used were: `n_estimators` = 1000, `max_depth` = 3, `learning_rate` = 0.05, `subsample` = 0.8, `colsample_bytree` = 0.8, `reg_alpha` = 0.1, `reg_lambda` = 0.1, `random_state` = 42, and `early_stopping_rounds` = 20.

5.2.7 Performance Metrics

Model performance was assessed using three metrics:

- **Mean Absolute Error (MAE):**

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5.1}$$

- **Root Mean Squared Error (RMSE):**

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5.2}$$

- **Mean Absolute Relative Difference (MARD):**

$$\text{MARD} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \times 100\% \tag{5.3}$$

where y_i are the observed glucose values and $\hat{y}_i$ are the predicted values. All metrics were averaged across 5 cross-validation folds.

5.2.8 Clinical Error Grid Analysis

Clinical accuracy was assessed using two widely used measures: the Clarke Error Grid (CEG) and the Parkes Error Grid (PEG). Both classify predictions into zones based on the percentage difference from a reference value, but they differ in structure and clinical emphasis. The Parkes Error Grid, being more recent, places greater focus on the clinical consequences of hypoglycemia and hyperglycemia and features nonlinear boundaries to better capture the risks associated with these extremes (Rodríguez-Rodríguez et al., 2024). Zone A represents clinically accurate predictions within 20% of the reference value; Zone B represents benign errors with no significant clinical consequences; Zone C includes errors that could lead to unnecessary corrective actions; Zone D consists of potentially dangerous failures to detect hypo- or hyperglycemia; and Zone E contains erroneous predictions that could lead to treatment in the wrong direction.

5.3 Results

5.3.1 Demographics

The study included 54 patients with T1D (Table 5.1). The cohort's median age was 26 years, with a range of 19 to 70 and an IQR of 19 years. The 95% confidence interval (CI) for mean age was 28.2–36.3 years.

The population was largely female, comprising 37 patients (68.5%), a proportion significantly higher than the expected 50% ($p = 0.005$, binomial test). The 95% CI for the proportion of females was 56.1–80.9%.

In terms of racial demographics, the majority identified as White or Caucasian (88.9%), followed by Asian or Pacific Islander (3.7%), Black/African American (1.9%), Black/African American and White (1.9%), Other (1.9%), and Not Reported (1.9%).

Hemoglobin A1C test results were obtained for 51 patients (87.9%). The mean hemoglobin A1C level was 7.7%, with a standard deviation (SD) of 0.83% and a range of 6.3%–10%. The 95% CI for the mean was 7.543–8.010%. The mean was significantly higher than the healthy reference value of <5.7% ($p < 0.001$, t-test).

Table 5.1: Demographic characteristics of the Type 1 Diabetes Study Cohort ($N = 54$).

Characteristic	Value(s)	Difference	p-value
Age (years)			
Median	26	N/A	N/A
Range	19–70		
IQR	19		
95% CI	28.2–36.3		
Female sex ($\hat{p} = 50\%$)			
n (%)	37 (68.5)	18.50%	0.005 (binomial)
95% CI, %	56.1–80.9		
Race, n (%)			
White/Caucasian	48 (88.9)		
Asian or Pacific Islander	2 (3.7)		
Black/African American	1 (1.9)		
Black/African American & White	1 (1.9)		
Other	1 (1.9)		
Not reported	1 (1.9)		
Haemoglobin A1C			
n (%) tested	51 (87.9)		
Mean ($\mu_0 = 5.7$)	7.7	2	<0.001 (t-test)
Range	6.3–10		
SD	0.83		
95% CI	7.543–8.010		

5.3.2 Baseline Model Performance

Random Forest

The Random Forest model had possible hyperparameters: `n_estimators`: 100, 200, 300; `max_depth`: None, 10, 20; `min_samples_split`: 2, 5, 10; and `min_samples_leaf`: 1, 2, 4. Through 3 iterations of the Grey Wolf Optimizer, the optimal hyperparameter

set was found to be `n_estimators` = 100, `max_depth` = None, `min_samples_split` = 2, and `min_samples_leaf` = 1. The fitness value improved after each iteration, starting from 51 in Iteration 0 and ending at 42, indicating that the final set of hyperparameters outperforms the default.

Five-fold cross-validation yielded a cross-validated MAE of 16.07, RMSE of 23.04, and MARD of 9.97.

Long Short-Term Memory

The LSTM model had possible hyperparameters: `n_units`: 50, 100, 150; `learning_rate`: 0.001, 0.01, 0.1; `batch_size`: 16, 32, 64; and `epochs`: 50, 100. Through 3 iterations of the Grey Wolf Optimizer, the optimal set was found to be `n_units` = 50, `learning_rate` = 0.001, `batch_size` = 16, and `epochs` = 50.

Five-fold cross-validation yielded a cross-validated MAE of 13.28, RMSE of 18.95, and MARD of 8.24.

Gated Recurrent Unit

The GRU model had the same hyperparameter search space as LSTM. Through 3 iterations of the Grey Wolf Optimizer, the optimal set was also found to be `n_units` = 50, `learning_rate` = 0.001, `batch_size` = 16, and `epochs` = 50.

Five-fold cross-validation yielded a cross-validated MAE of 13.31, RMSE of 18.77, and MARD of 8.35.

These individual learners were selected for their complementary strengths. Random Forest excels at handling lagged variables and capturing short-term patterns, while LSTM and GRU are well-suited for modelling long-term dependencies in time series data. In the upward-trending portions of the prediction plots, LSTM tended to overpredict while GRU tended to underpredict, indicating that neither model performs optimally on its own across all scenarios. By combining their outputs, the meta-learner

can leverage the strengths of each model to produce a more accurate and robust final prediction.

5.3.3 Stacking Ensemble Performance (XGBoost)

Averaging across 5 folds, the XGBoost stacking ensemble meta-learner yielded an MAE of 10.65, RMSE of 14.59, and MARD of 6.97, outperforming all three baseline models across all metrics.

5.3.4 Comparison of Overall Performance

Table 5.2 presents the full performance comparison across all four models. The XGBoost Stacked Model consistently outperformed the baseline models, with the lowest MAE (10.65), RMSE (14.59), and MARD (6.98). Among the baseline models, GRU performed best, with MAE (13.31), RMSE (18.77), and MARD (8.35), followed closely by LSTM with MAE (13.28), RMSE (18.95), and MARD (8.24). Random Forest performed noticeably worse than GRU and LSTM across all three metrics.

Table 5.2: Comparison of model accuracy with Grey Wolf Optimization. Results averaged across 5-fold cross-validation.

Algorithm	Type	MAE	RMSE	MARD
Random Forest	Machine learning	16.07	23.04	9.97
GRU	Machine learning	13.31	18.77	8.24
LSTM	Machine learning	13.28	18.95	8.35
XGBoost	Stacking ensemble	10.65	14.59	6.98

5.3.5 Grey Wolf as a Requirement for Optimal Performance

The inclusion of GWO was found to be critical for reaching state-of-the-art performance. Without Grey Wolf, the Random Forest, LSTM, and GRU methods did not perform as well as with optimised hyperparameters. The XGBoost algorithm likewise

underperformed without Grey Wolf, exhibiting much higher error rates across all measures. These findings emphasise the relevance of GWO in enabling the baseline learners to work optimally. Some algorithms showed decent performance without optimisation, possibly due to a fortuitous default choice of hyperparameter values; however, the XGBoost stacked model with GWO emerged as the best overall model.

5.3.6 Clinical Error Grid Analysis

A high proportion of the XGBoost meta-learner's predictions - over 90% - fell within Zone A of the Clarke Error Grid, indicating predictions within 20% of reference values and thus clinically accurate. Table 5.3 shows the distribution of predictions across CEG zones for each fold. In each fold, 89%–97% of predictions fell within Zone A, with smaller percentages in Zone B (3%–8%), Zone C (1%–2%), and very few in Zones D or E (0%–2%).

Table 5.3: Clarke Error Grid zone distribution for the XGBoost meta-learner across 5 cross-validation folds.

Fold	Zone A	Zone B	Zone C	Zone D	Zone E
1	91.20%	5.41%	1.02%	0.55%	1.83%
2	94.32%	4.93%	0.50%	0.20%	0.05%
3	88.92%	8.21%	1.86%	0.87%	0.45%
4	95.69%	3.79%	0.30%	0.22%	0.00%
5	96.08%	3.25%	0.42%	0.20%	0.05%

The Parkes Error Grid analysis of the same model found that the majority of observations fell within Zone A, with a smaller number in Zone B representing minor, acceptable deviations. Notably, no predictions fell within Zones C, D, or E, indicating the model avoids potentially dangerous errors and further supporting its suitability for clinical use (Rodríguez-Rodríguez et al., 2024).

5.4 Discussion

5.4.1 Summary of Findings

Using data from 54 patients with T1D, this study created and evaluated three machine learning baseline algorithms and one stacking ensemble meta-learner for predicting blood glucose levels 15 minutes ahead. GRU consistently outperformed the other baseline models, and the XGBoost stacked model outperformed each individual baseline model. Critically, the incorporation of GWO was found to be essential for the XGBoost algorithm to reach state-of-the-art performance. Without GWO, the baseline models underperformed with greater error rates across all metrics, contributing to less-than-ideal stacked model performance.

5.4.2 Comparison with Similar Research

Our findings demonstrate that GWO is critical for obtaining cutting-edge performance in forecasting blood glucose levels. A recent study by Yang et al. (2024) developed an enhanced stacking ensemble learning method that achieved world-class predictive performance with a 30-minute prediction horizon RMSE of 1.425 mg/dL and MAE of 0.721 mg/dL. Notably, that study used a 30-minute prediction horizon compared to our 15-minute window, which affects direct comparison of error metrics. The enhanced stacking ensemble in Yang et al. (2024) outperformed the non-ensemble StackLSTM model with RMSE and MAE gains of up to 27.92% and 65.32%, respectively, which is consistent with our finding that stacking improved on our individual baseline learners.

5.4.3 Hybrid Approaches

In this analysis, GWO was applied only to the machine learning algorithms, and under this scenario GRU was the most accurate baseline algorithm. Other studies have used various methods analogous to GWO to adaptively optimise ARIMA, GRU, LSTM,

and other time series and machine learning algorithms, achieving better results than standalone models (Alkanhel et al., 2024; Swain et al., 2022). One study reported an MSE of 0.330 using ARIMA based on Bayesian optimisation, which when converted to RMSE suggests state-of-the-art performance (QingXiang et al., 2023), though a potential caveat is that a shorter prediction time window may have led to lower error metrics. A hybrid Transformer-LSTM model has also demonstrated state-of-the-art RMSE at all time points (Bian et al., 2024), though without MAE or MAPE values being reported, it is unclear whether those results represent an improvement over the findings presented here.

The inclusion of bolus data did not significantly improve the predictive accuracy of the Random Forest model. This may be attributed to noise introduced by additional covariates - such as carbohydrate intake, insulin intake, and insulin on board - which, while clinically relevant, did not markedly affect model performance. These findings suggest that the value of incorporating detailed bolus and insulin pump data should be carefully evaluated, particularly given the cost and complexity of collecting such information.

5.4.4 Strengths and Limitations

Strengths

This study has several strengths. First, it comprised a well-characterised cohort of 54 T1D patients with over 27,000 days of CGM and over 8,000 days of insulin pump data, providing a rich dataset for comparing multiple blood glucose prediction algorithms. Second, the cohort's demographic characteristics - including age, gender, and race - were extensively characterised, increasing the generalisability of the results. Third, the inclusion of haemoglobin A1C data for 51 of the 54 patients provided useful information on cohort glycemic management. Fourth, the use of a diverse set of machine learning and time series models enabled comprehensive comparison

of model effectiveness. Finally, the identification of GWO as a vital component for optimal performance emphasised the importance of rigorous hyperparameter tuning and stacking ensemble methods.

Limitations

Several limitations should be acknowledged. First, the cohort's demographics, which include a higher proportion of females (68.5%), may introduce a small bias. However, it is not clear that gender is a significant confounding variable in blood glucose prediction; while hormonal and lifestyle differences may contribute to subtle variations in glucose dynamics, these effects are likely modest. Further subgroup analyses could clarify whether gender-specific patterns exist.

Second, the study examined only T1D, and the models' effectiveness may differ when applied to patients with type 2 diabetes.

Third, the study used discrete hyperparameter tuning, which may not have identified the absolutely optimal hyperparameter values. Because the parameter space for each model was pre-constructed from discrete values, superior parameter values between the predefined ones may have been missed. Nonetheless, the study demonstrates the versatility and utility of GWO in selecting the best hyperparameters from a set of reasonable candidate values.

Finally, the study did not account for potential population or disease changes over time, which may impact model performance. Future research should employ more varied and representative datasets, investigate other machine learning techniques, and examine different prediction horizons beyond the 15-minute window used here.

5.5 Conclusion

This study developed and evaluated machine learning models to construct a highly accurate blood glucose prediction method for patients with T1D, using the DiaTrend dataset. The GRU technique consistently outperformed the other baseline models, achieving the lowest MAE, RMSE, and MARD among the baseline learners. The Stacking Ensemble XGBoost meta-learner further improved predictive accuracy across all metrics. Critically, the use of the Grey Wolf Optimizer was essential to the baseline models' and XGBoost meta-learner's optimal performance.

These findings demonstrate that machine learning algorithms, including stacking ensemble approaches, can reliably predict blood glucose levels, with important implications for T1D management. By incorporating these predictive algorithms into CGM devices, patients can be forewarned about impending blood glucose abnormalities, allowing appropriate and timely diabetes self-management to reduce or prevent symptoms, complications, and hospitalisations.

Chapter 6

Optimizing Sepsis Treatment via Latent Protocol Inference

6.1 Introduction

6.1.1 Clinical Motivation

Sepsis, defined as life-threatening organ dysfunction caused by a dysregulated host response to infection (Singer et al., 2016), accounts for at least 1.7 million adult hospitalisations and 350,000 deaths annually in the United States (Dantes, 2023). Despite its prevalence, treatment guidelines for high-stakes interventions - such as intravenous fluids and vasopressors - remain inconclusive, leading to highly variable practice patterns and suboptimal outcomes (Byrne and Van Haren, 2017). This lack of personalised decision support indicates a critical need for unified, data-driven tools (Komorowski et al., 2018).

The sequential nature of sepsis care naturally fits the Markov Decision Process (MDP) framework. Patient trajectories form state-action sequences where outcomes (survival) are observed only at the end of the episode. While reinforcement learning (RL) has been widely applied to this domain (Raghu et al., 2017; Komorowski et al.,

2018; Peng et al., 2018; Choi et al., 2024), standard methods face significant challenges in retrospective settings.

6.1.2 Sequence Modeling Approaches

Recent work has shifted towards offline, trajectory-centric learning, framing healthcare as sequence modelling (Liventsev and Fritz, 2024). Algorithms such as Diffusion Policy (Janner et al., 2022) and Decision Transformers (DT) (Chen et al., 2021) have shown promise by modelling trajectories directly. However, in the context of sepsis, these methods struggle with two key limitations:

1. **Sparse, delayed rewards:** Binary survival outcomes make credit assignment difficult for return-conditioned models.

2. **Lack of temporal abstraction:** Flat policies often fail to capture the hierarchical, protocol-level structure of clinical care.

6.1.3 Generative Latent Protocol Model

This chapter proposes a generative modeling perspective. We hypothesise that effective clinical strategies lie on a low-dimensional manifold embedded within the high-dimensional space of patient trajectories. We capture this structure using a latent variable z, representing the underlying *treatment protocol*. Instead of learning a fixed policy mapping states to actions, we learn a joint distribution over protocols, trajectories, and outcomes.

This probabilistic formulation enables a robust mechanism for personalised treatment planning that operates on two distinct timescales. First, we perform **global learning** to capture general medical knowledge and physiological dynamics - by training on the entire population via Maximum Likelihood Estimation (MLE), the model learns general laws of the environment. Second, we enable **test-time adaptation** for

individual cases: rather than applying a fixed policy, we treat the latent protocol z as a tunable parameter and perform an inference step to identify the specific protocol that maximises the probability of survival for a new patient.

6.1.4 Contributions

The main contributions of this chapter are:

1. **Generative Latent Protocol Model (GLP):** A hierarchical sequence model that captures treatment strategies via a latent bottleneck, trained using an Expectation-Maximisation (EM) framework on trajectory-outcome pairs.

2. **Goal-Conditioned Posterior Inference:** Treatment recommendation is formulated as a probabilistic inference problem; sampling from the survival-conditioned posterior allows for precise control over treatment aggressiveness.

3. **State-of-the-Art Performance:** On the ICU-Sepsis benchmark derived from MIMIC-III (Johnson et al., 2016), the GLP model consistently outperforms offline baselines across datasets of varying quality.

4. **Interpretability:** The learned latent space captures clinically meaningful treatment phenotypes (e.g., "Wet" vs. "Dry" fluid strategies) and enables counterfactual analysis to identify missed interventions in non-survivor trajectories.

6.2 Methods

6.2.1 Problem Formulation

Sepsis treatment is formulated as a finite-horizon sequential decision-making problem. A patient trajectory is denoted by $\tau = (s_0, a_0, s_1, a_1, \ldots, s_H, a_H)$, where $s_t \in \mathcal{S}$ is the physiological state, $a_t \in \mathcal{A}$ is the clinical intervention, and H is the horizon. Each

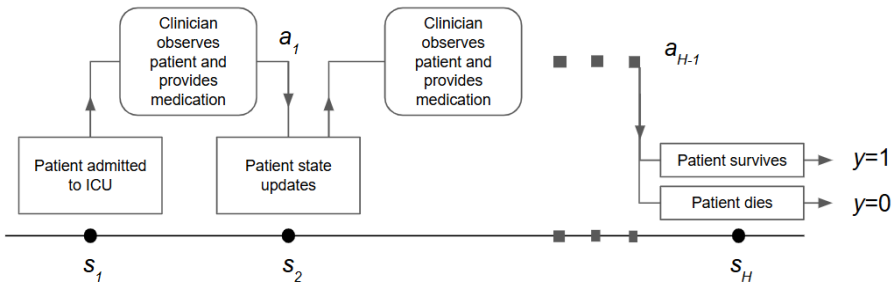

Figure 6.1: Illustration of a patient trajectory in the sequential decision-making framework.

trajectory is associated with a binary terminal outcome $y \in \{0, 1\}$ indicating patient survival (Figure 6.1).

Conventional RL approaches optimise a policy $\pi(a_t|s_t)$ to maximise expected return. However, sepsis treatment data is characterised by sparse, delayed rewards (outcome observed only at $t = H$) and complex physiological dynamics, making credit assignment difficult for standard value-based methods. Instead of learning a policy directly - and instead of conditioning on Return-to-Go (RTG) as in the Decision Transformer (Chen et al., 2021), which is a noisy signal in binary-outcome settings - this chapter adopts a generative sequence modelling perspective, explicitly modelling the underlying treatment strategy via a latent variable.

6.2.2 Generative Latent Protocol Model

We propose a generative model that treats the treatment policy and patient outcome as conditionally independent given a latent *protocol* variable $z \in \mathbb{R}^d$. This latent variable acts as an information bottleneck, capturing the high-level medical strategy governing the episode (Figure 6.2). The joint distribution is defined as:

$$p_\theta(\tau, y, z) = p_\alpha(z)\, p_\beta(\tau|z)\, p_\gamma(y|z), \tag{6.1}$$

where $\theta = (\alpha, \beta, \gamma)$ are learnable parameters. The generative process has three

components:

Latent Protocol Prior

The protocol z is generated by a neural transformation U_α acting on Gaussian noise:

$$z = U_\alpha(z_0), \quad z_0 \sim \mathcal{N}(0, I_d). \tag{6.2}$$

U_α is parameterised by a non-linear network (a UNet (Ronneberger et al., 2015)), allowing the model to learn a complex manifold of valid treatment protocols from a simple base distribution.

Treatment Trajectory Generator

Conditioned on the protocol z, the trajectory τ is generated autoregressively via a causal Transformer:

$$p_\beta(\tau|z) = \prod_{t=1}^{H} p_\beta(s_t, a_t \mid s_0, a_0, \ldots, s_{t-1}, a_{t-1}, z), \tag{6.3}$$

where the protocol z is injected via cross-attention layers, ensuring temporal consistency across the treatment episode.

Outcome Predictor

The relationship between a protocol and patient survival is modelled by a discriminator $f_\gamma : \mathbb{R}^d \to [0, 1]$, using a Bernoulli likelihood with $f_\gamma(z) = \sigma(s_\gamma(z))$:

$$p_\gamma(y|z) = (f_\gamma(z))^y (1 - f_\gamma(z))^{1-y}. \tag{6.4}$$

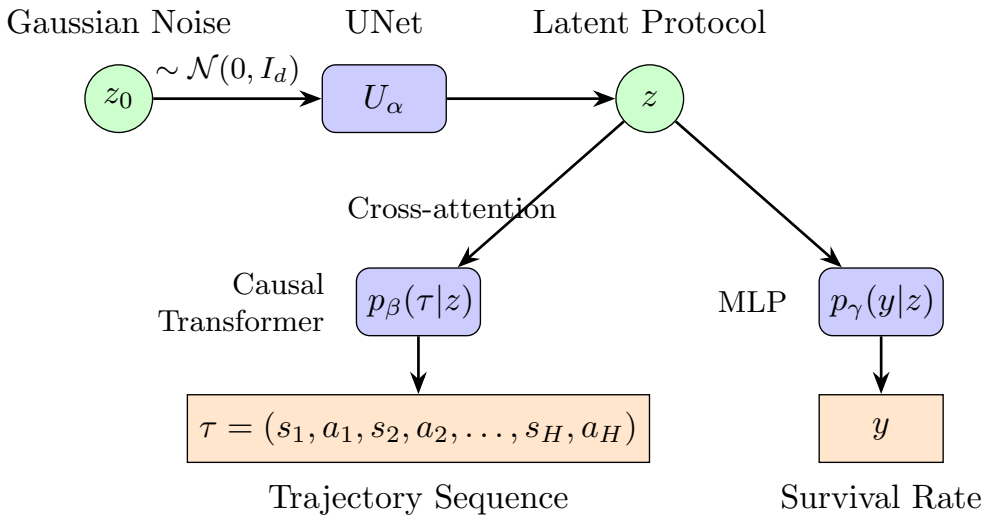

Figure 6.2: Overview of the proposed framework.

6.2.3 Maximum Likelihood Learning

We learn the model parameters θ by maximising the marginal log-likelihood of the observed offline dataset $\mathcal{D} = \{(\tau_i, y_i)\}_{i=1}^n$:

$$L(\theta) = \log \int p_\beta(\tau|U_\alpha(z_0))\, p_\gamma(y|U_\alpha(z_0))\, p_0(z_0)\, dz_0. \tag{6.5}$$

Using the Fisher identity, the gradient can be expressed as an expectation over the posterior:

$$\nabla_\theta L(\theta) = \mathbb{E}_{z_0 \sim p_\theta(z_0|\tau,y)} \left[\nabla_\theta \log p_\theta(\tau, y, U_\alpha(z_0))\right]. \tag{6.6}$$

This formulation naturally suggests an iterative EM learning scheme:

- **E-Step (Inference):** For each trajectory-outcome pair (τ, y), infer the posterior $p_\theta(z_0|\tau, y)$, identifying the latent protocol that best explains the observed decisions and outcome.
- **M-Step (Learning):** Update parameters θ via stochastic gradient ascent to maximise the joint likelihood given the inferred protocols.

The posterior distribution is characterised by an energy function:

$$E(z_0) = \frac{1}{2}\|z_0\|^2 - \log p_\beta(\tau|U_\alpha(z_0)) - \log p_\gamma(y|U_\alpha(z_0)), \tag{6.7}$$

from which we sample using the approximate posterior sampling strategies described in Section 6.2.5.

6.2.4 Fast-Slow Adaptation Mechanism

The EM framework admits an interpretation in terms of two timescales of parameter adaptation:

- **Global "Slow" Parameters (θ):** Network weights shared across the entire population, capturing the general laws of physiology and medical best practices, learned slowly over the full dataset.
- **Local "Fast" Parameters (z):** The latent protocol acting as a patient-specific parameter vector, optimised via posterior sampling during the E-step for the current episode.

This separation enables single-instance adaptation: the model effectively tailors the policy to a specific patient without modifying the global network weights, retaining the robustness of a population-level model.

6.2.5 Posterior Inference Strategies

Three strategies are investigated for approximating the posterior expectation, each offering different trade-offs between mixing accuracy, computational cost, and implementation complexity.

Short-Run MCMC (Overdamped Langevin)

We approximate the posterior using overdamped Langevin dynamics with a short-run scheme (Nijkamp et al., 2019). Chains are initialised from the prior and run for K steps:

$$z_0^{(k+1)} = z_0^{(k)} - \frac{s}{2}\nabla_{z_0} E(z_0^{(k)}) + \sqrt{s}\,\epsilon^{(k)}, \quad \epsilon^{(k)} \sim \mathcal{N}(0, I). \tag{6.8}$$

Underdamped Langevin Dynamics (BAOAB)

To accelerate convergence, we employ underdamped Langevin dynamics augmented with a momentum variable v. The system is solved numerically using the BAOAB integrator (Leimkuhler and Matthews, 2013), which applies Force (B), Drift (A), and Ornstein-Uhlenbeck (O) steps symmetrically:

$$\begin{aligned}
&\text{B:} \quad v \leftarrow v - \frac{\eta}{2}\nabla_{z_0} E(z_0) \\
&\text{A:} \quad z_0 \leftarrow z_0 + \frac{\eta}{2} v \\
&\text{O:} \quad v \leftarrow e^{-\gamma\eta} v + \sqrt{1 - e^{-2\gamma\eta}}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \\
&\text{A:} \quad z_0 \leftarrow z_0 + \frac{\eta}{2} v \\
&\text{B:} \quad v \leftarrow v - \frac{\eta}{2}\nabla_{z_0} E(z_0)
\end{aligned}$$

The symmetric composition ensures time-reversibility and $\mathcal{O}(\eta^3)$ local error.

Stochastic Variational Inference (SVI)

As an alternative to MCMC, we approximate the posterior with a diagonal Gaussian $q_\phi(z_0) = \mathcal{N}(\mu, \mathrm{diag}(\sigma^2))$, directly optimising variational parameters $\phi = \{\mu, \sigma\}$ for each episode to maximise the ELBO:

$$\mathcal{L}_{\mathrm{ELBO}}(\phi) = \mathbb{E}_{q_\phi(z_0)}\left[\log p_\beta(\tau | U_\alpha(z_0)) + \log p_\gamma(y | U_\alpha(z_0))\right] - D_{\mathrm{KL}}\left(q_\phi(z_0) \| p_0(z_0)\right). \tag{6.9}$$

Gradients are estimated via the reparameterisation trick (Kingma and Welling, 2013).

6.2.6 Protocol-Conditioned Treatment at Inference

Once trained, the model functions as a goal-conditioned policy. Treatment trajectories can be generated conditioned on a target survival probability $y^\star \in [0, 1]$ via:

$$p(\tau|y^\star) = \int p(z_0|y^\star)\, p_\beta(\tau|z = U_\alpha(z_0))\, dz_0. \tag{6.10}$$

Stage 1: Latent Protocol Planning

We infer the latent protocol by sampling from the target-conditioned posterior:

$$p(z_0|y^\star) \propto p_0(z_0)\, p_\gamma(y^\star|z = U_\alpha(z_0)), \tag{6.11}$$

using the planning energy $E_{\text{plan}}(z_0) = \frac{1}{2}\|z_0\|^2 - \log p_\gamma(y^\star|U_\alpha(z_0))$.

The target $y^\star$ is interpreted as a continuous risk profile using binary cross-entropy: $-\log p_\gamma(y^\star|z) \coloneqq \text{CE}(y^\star, f_\gamma(z))$. The gradient of the planning energy is:

$$\nabla_{z_0} E_{\text{plan}}(z_0) = z_0 - \frac{1}{T_y}(y^\star - f_\gamma(z))\nabla_{z_0} s_\gamma(z), \tag{6.12}$$

where T_y acts as a temperature scaling factor. The resulting sample $z^\star$ represents an optimised treatment plan.

Stage 2: Action Execution

Given the inferred protocol $z^\star$, actions are generated autoregressively at each timestep:

$$a_t \sim \pi_\beta(a_t|s_{\leq t}, a_{<t}, z^\star). \tag{6.13}$$

Keeping $z^\star$ fixed throughout the episode ensures that individual decisions remain temporally coherent and aligned with the high-level survival strategy identified during planning.

6.2.7 Implementation Details

The latent protocol prior U_α is parameterised by a 1D U-Net with latent dimension $d = 128$. The trajectory generator p_β employs a 2-layer causal Transformer with 4 attention heads, an embedding dimension of 128, and a context window of $K = 10$. The outcome predictor p_γ is a 3-layer MLP with hidden units [64, 64]. Training used the Adam optimiser (Kingma and Ba, 2014) with learning rate 1×10^{-4} and batch size 64. For posterior sampling in the E-step, $N = 5$ steps were used during training; at inference time this was increased to $N = 25$ for more accurate latent protocol estimation. Training was performed on a single NVIDIA A6000 GPU, converging in approximately 5 hours.

6.3 Data

6.3.1 ICU-Sepsis Benchmark

To evaluate the proposed method, we utilised the ICU-Sepsis environment (Choudhary et al., 2024), a standardised tabular MDP modeling sepsis management. Constructed from over 40,000 patient records in the MIMIC-III database (Johnson et al., 2016), the environment poses a challenging sequential decision-making task with 716 discrete physiological states and a sparse, binary survival outcome. The action space consists of 25 discrete treatment combinations, representing discretised dosages of intravenous (IV) fluids and vasopressors (Table 6.1).

Table 6.1: Action space discretization. IV fluids and vasopressors are each binned into five dosage levels, yielding $5 \times 5 = 25$ treatment combinations.

Level	IV Fluids (mL/hr)	Vasopressor (μg/kg/min)
0	0	0
1	0–49	0–0.08
2	50–149	0.08–0.20
3	150–499	0.20–0.45
4	>499	>0.45

6.3.2 Offline Dataset Construction

Following the D4RL methodology (Fu et al., 2020), five offline datasets of varying quality were constructed by training Soft Actor-Critic (SAC) agents online and capturing checkpoints at different performance thresholds:

- **Random:** 1M steps collected by a randomly initialised policy.
- **Medium-Replay:** The full replay buffer of a policy trained to medium-level performance.
- **Medium:** 1M steps from a fixed, partially-trained policy.
- **Expert:** 1M steps from a fully-converged SAC policy (approximately 83% survival rate).
- **Real:** Real-world clinical trajectories from MIMIC-III, mapped into the ICU-Sepsis MDP format following Komorowski et al. (2018).

6.4 Results

6.4.1 Main Results

Table 6.2 presents the comparative performance of GLP against four baselines: Behavior Cloning (BC) (Atkeson and Schaal, 1997), Conservative Q-Learning (CQL)

(Kumar et al., 2020), Decision Transformer (DT) (Chen et al., 2021), and Q-Learning Decision Transformer (QDT) (Yamagata et al., 2023). All experiments were repeated across 5 random seeds; results are reported as mean ± standard deviation survival rate (%).

Table 6.2: Performance on ICU-Sepsis (mean ± std survival rate, %). Our generative approach consistently outperforms baselines. Bold indicates the best result in each column.

Algorithm	Random	Med.-Replay	Medium	Expert	Real
BC	79.80±3.66	80.00±4.34	84.00±2.68	88.00±3.29	79.40±3.44
CQL	76.60±3.72	80.60±4.22	79.20±3.97	77.60±2.42	78.40±3.55
DT	78.20±3.25	79.20±2.14	84.00±2.97	87.40±3.93	80.20±2.93
QDT	80.80±4.17	83.00±1.67	86.00±1.10	88.00±2.61	80.40±4.13
GLP (Short-Run)	**81.60±3.01**	87.00±3.36	87.60±2.42	91.00±2.41	80.20±1.17
GLP (BAOAB)	80.50±2.77	**89.00±3.13**	**89.00±4.00**	**91.00±2.17**	**83.20±2.68**
GLP (SVI)	81.60±6.02	85.00±2.68	87.00±1.79	88.00±2.10	82.00±4.47

GLP consistently outperforms baselines across diverse data qualities. On the Medium-Replay and Medium datasets, the BAOAB variant achieves survival rates of 89.0%, substantially surpassing the best baseline (QDT: 86.0%) and the behaviour policies used to collect the data. On the Expert dataset, both Short-Run MCMC and BAOAB break the 88.0% plateau observed in BC and QDT, achieving 91.0%. On the Real dataset - the hardest challenge due to noise and confounding - BAOAB achieves 83.2%, meaningfully improving on the approximately 80% performance of standard offline RL baselines. Overall, Short-Run MCMC provides strong performance on lower-quality data (Random), while the superior mixing properties of BAOAB yield the most robust policies as dataset quality improves.

6.4.2 Optimal Target Selection

We conducted an ablation study evaluating the sensitivity of GLP to different inference targets $y^\star \in \{0.85, 0.90, 0.95, 1.00\}$ (Table 6.3). Moderate targets ($y^\star \in \{0.90, 0.95\}$) proved most effective, with $y^\star = 0.95$ consistently yielding the best performance.

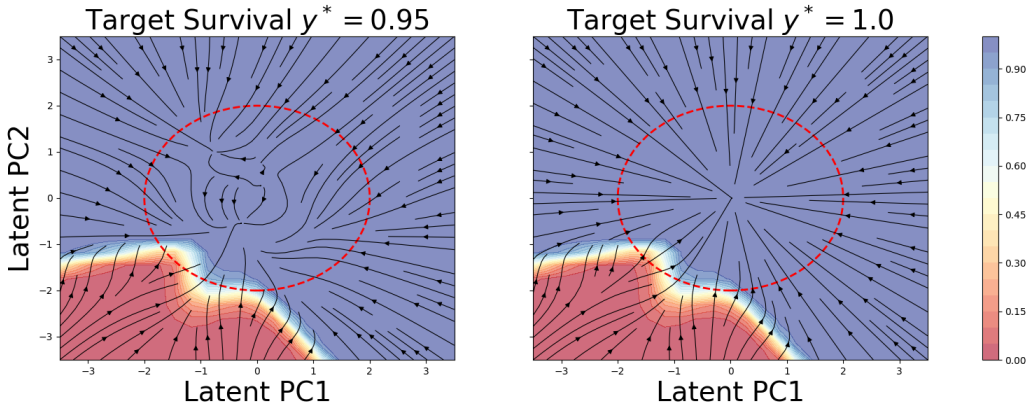

Figure 6.3: Survival sensitivity analysis with respect to the latent policy temperature.

Pushing the target to theoretical perfection ($y^\star = 1.00$) degraded performance, suggesting that $y^\star = 0.95$ finds a high-likelihood, robust protocol, whereas $y^\star = 1.00$ pushes the latent code into out-of-distribution regions.

Table 6.3: Ablation study of performance across target survival rates (mean ± std, %). Results averaged over 5 seeds; bold indicates best result per dataset.

Dataset	$y^\star = 0.85$	$y^\star = 0.90$	$y^\star = 0.95$	$y^\star = 1.00$	Intervention level
Random	80.00±5.10	77.80±3.06	**81.60±3.01**	79.60±4.03	Conservative
Med.-Replay	86.20±2.32	87.40±3.44	**87.00±3.36**	84.80±2.32	Balanced
Medium	82.30±1.83	84.00±2.97	**87.60±2.42**	85.80±3.49	Balanced
Expert	86.40±3.98	85.40±4.59	**91.00±2.41**	88.00±3.16	Balanced
Real	77.20±2.32	77.40±2.73	**80.20±1.17**	76.60±2.87	Conservative

Visualisation of the gradient field $\nabla_z \log p(z|y^\star)$ reveals a *gradient conflict* driven by a prior-likelihood mismatch at extreme targets (Figure 6.3). For $y^\star = 1.00$, the vector field becomes nearly radial and is dominated by contraction toward a single attractor. In contrast, $y^\star = 0.95$ creates a balanced vector field that navigates z toward high-survival regions while staying anchored to the prior support, confirming that the model performs best when searching for the best *realistic* protocol rather than an idealised outlier.

6.4.3 Qualitative Analysis and Interpretability

Latent Space Geometry

To verify whether the latent space captures coherent clinical strategies, we linearly interpolated between learned "Wet" (high-fluid) and "Dry" (low-fluid) protocol centroids. Varying the interpolation factor α primarily modulated fluid recommendations, inducing earlier and more pronounced fluid administration, while the vasopressor trajectory remained comparatively robust. This suggests the model has disentangled the timing of fluid resuscitation from vasopressor support within the latent representation.

Counterfactual Analysis

We analysed non-survivor trajectories to identify specific missed opportunities. The model was conditioned on $y^\star = 0.95$ and the history of Patient 27 was replayed (Figure 6.4). The model diverged from the clinician at 21% of time steps (7 of 33 steps), revealing two key corrections:

1. **Early fluid intervention:** At $t = 1$ rather than the clinician's delayed bolus at $t = 2$.

2. **Prevention of premature vasopressor weaning:** The model maintained steady vasopressor support during critical windows ($t \in \{10, 11, 20, 21\}$) while the clinician de-escalated.

The model effectively corrected the reactive "rollercoaster" pattern of care, favouring proactive stability. Across 50 non-survivors, this pattern generalised: GLP favoured sustained vasopressor support ($+1.3 \pm 4.5$ levels net change) while maintaining context-dependent fluid management ($+0.1 \pm 2.9$ levels net change), with de-escalation prevention occurring on average 1.8 ± 1.5 times per patient—the primary mechanism of divergence (Table 6.4).

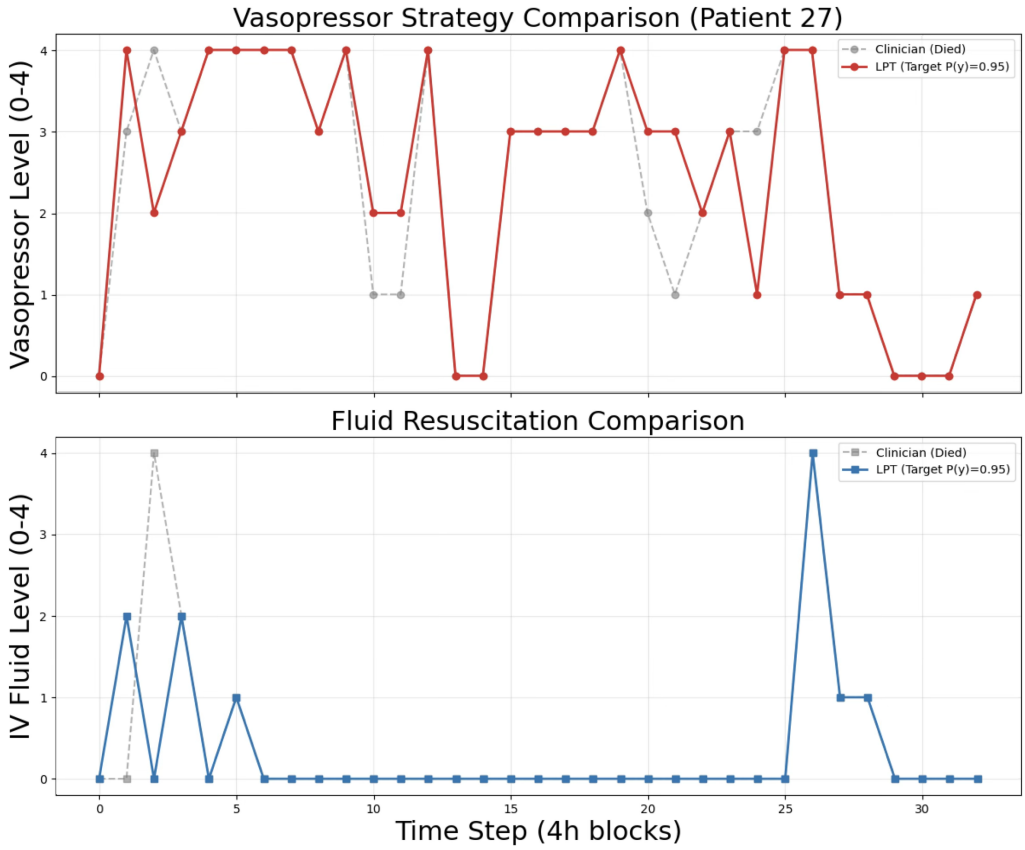

Figure 6.4: Counterfactual analysis for Patient 27.

Population-Level Action Distributions

Compared to observed clinician behavior, GLP consistently shifted towards higher vasopressor usage and restricted fluid administration (Figure 6.5). This identifies an alternative, more aggressive hemodynamic management strategy that correlates with higher survival within this environment.

6.5 Discussion

6.5.1 Summary of Findings

Using the ICU-Sepsis benchmark, this chapter created and evaluated a protocol-conditioned generative model for sepsis treatment optimisation that consistently

Table 6.4: Counterfactual analysis summary across $N = 50$ non-survivors.

Metric	Mean ± SD	Interpretation
Divergence rate	24.2% ± 9.8%	Selective intervention
Net vasopressor change	$+1.3 \pm 4.5$ levels	Sustained haemodynamic support
Net fluid change	$+0.1 \pm 2.9$ levels	Context-dependent fluid strategy
De-escalation preventions	1.8 ± 1.5 per patient	Primary divergence mechanism

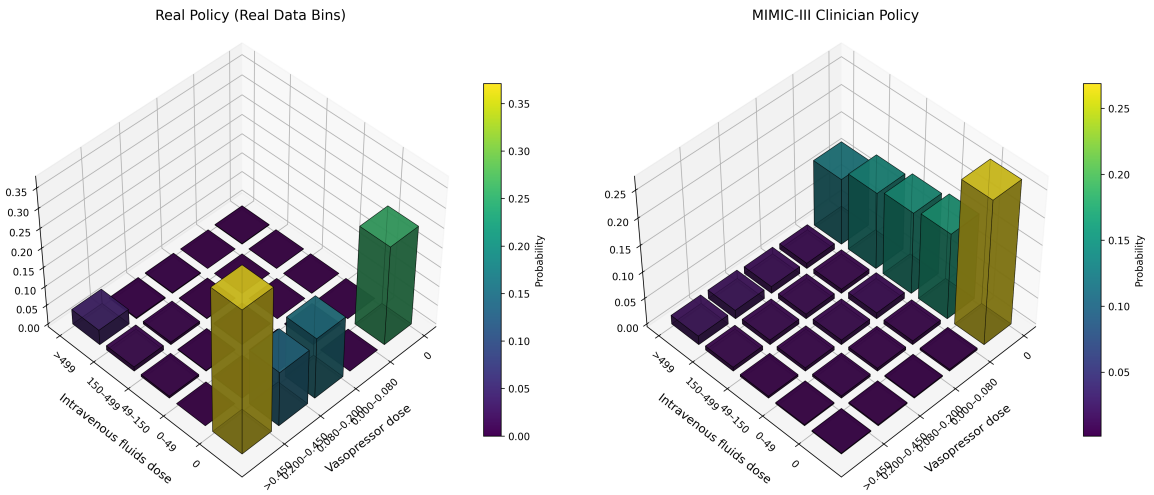

Figure 6.5: Action distribution comparison between the learned policy and the real MIMIC-III policy.

outperforms established offline RL baselines. A key advantage of the approach is the ability to perform goal-conditioned posterior inference over a latent protocol, enabling iterative refinement of treatment strategies toward a desired survival target.

6.5.2 Mechanism and Advantages

The method's performance stems from two design choices: learning from full trajectory-outcome pairs via MLE, and using a latent variable z for temporal abstraction. Unlike Behavior Cloning, which suffers from compounding errors over long horizons (Ross et al., 2011), the generative approach improves stability by guiding action generation via the latent protocol. Unlike CQL or DT variants, which can struggle with sparse rewards or noisy return-to-go targets (Kumar et al., 2020; Chen et al., 2021), GLP avoids value estimation biases by directly optimising the joint distribution $p_\theta(\tau, y, z)$.

This separates protocol planning from action generation, offering a robust alternative to value-based offline RL.

6.5.3 Clinician vs. AI Dosing

Qualitatively, the learned policies tended to recommend higher vasopressor doses and lower IV fluid volumes than clinician behaviour. While consistent with prior observations where AI clinicians recommend increased vasopressor usage (Komorowski et al., 2018), these trends reflect the underlying dynamics of the simulation benchmark and require careful validation before real-world interpretation.

6.5.4 Limitations

Several limitations should be acknowledged. First, the ICU-Sepsis environment is a simplified tabular MDP with discrete states and a limited action space, preventing it from capturing the full physiological complexity of sepsis care or intermediate outcomes such as organ recovery. Second, counterfactual trajectories were generated within the simulation - replaying the patient's *observed* state sequence under the counterfactual protocol. In reality, different actions would lead to different subsequent states, so these results represent what GLP would have recommended rather than what would have happened. Third, the counterfactual analysis examined only non-survivors with sufficiently long trajectories ($\geq$ 20 time steps), and patterns may not generalise to shorter stays or survivors. Fourth, the analysis does not establish causal efficacy; prospective validation would be required to determine whether GLP's recommendations actually improve outcomes.

6.5.5 Future Directions

Future work will extend latent protocol inference to real-world cohorts with irregular sampling and confounding, such as MIMIC-IV (Huang et al., 2025), and focus on integrating these protocols into clinical decision support workflows (Sendak et al., 2020).

6.6 Conclusion

We introduced a latent protocol inference framework that provides temporal abstraction and enables goal-conditioned planning without reliance on step-wise rewards. On the ICU-Sepsis benchmark, the GLP model consistently outperforms offline RL baselines including BC, CQL, DT, and QDT across datasets of varying quality. The BAOAB posterior sampling variant yielded the most robust performance overall. Representing treatment strategies as latent vectors allows for superior performance while enabling clinically meaningful interpretability through latent space geometry and counterfactual analysis.

Importantly, in the near term these findings should not be used to inform medical practice. The Generative Latent Protocol model is a methodological contribution that could inform future research on decision support for sepsis care, but any clinical application would require extensive prospective evaluation using real-time data and clinical trials.

Chapter 7

An Explainable Machine Learning Model for Hypertension Screening in Type 1 Diabetes Patients Using Continuous Glucose Monitoring Data

7.1 Introduction

Type 1 diabetes (T1D) affects approximately 1.6 million Americans and requires lifelong insulin therapy and intensive glycaemic monitoring. Continuous glucose monitoring (CGM) has become the standard of care, with devices worn continuously to track interstitial glucose levels every 5-15 minutes (Danne et al., 2017), generating rich, longitudinal data. T1D patients are at elevated risk for multiple comorbidities, including hypertension, thyroid disorders, retinopathy, and dyslipidemia (De Ferranti et al., 2014). Hypertension affects approximately 10-30% of T1D patients and substantially

increases cardiovascular disease risk (Thorn et al., 2005), while hypothyroidism affects 17-30% of T1D patients (Kadiyala et al., 2010).

CGM technology generates rich longitudinal data capturing not only average glucose levels but also variability, hypoglycemic events, and temporal patterns, while HbA1c is a measure of glycemic control for the past three months. Studies have shown that CGM data can predict glucose trends and hypoglycemic events (Liu et al., 2025), and that glycemic variability correlates with cardiovascular outcomes (Ceriello et al., 2008). Prior research has primarily focused on using CGM for glucose prediction rather than comorbidity risk classification. However, the utility of CGM data for classifying other outcomes is starting to be studied, with recent work showing that CGM metrics such as time in range (TIR) and glycaemic variability were associated with all-cause mortality to a greater extent than HbA1c (Okuno et al., 2025). Recent advances in machine learning have demonstrated the potential to extract clinical insights from wearable device data beyond their primary purpose (Okuno et al., 2025).

7.1.1 Biological Rationale for CGM-Hypertension Linkage

We hypothesised that CGM metrics would provide added discrimination value specifically for diagnosed hypertension classification. This is based on the principle that including features which provide additional discrimination information improves models, while adding redundant features does not (Guyon and Elisseeff, 2003). The biological plausibility of a link between T1D and hypertension stems from established connections between glycaemic variability and endothelial dysfunction (Monnier et al., 2006), and between hypoglycaemia and autonomic dysregulation (Reno et al., 2013), both of which may influence blood pressure regulation.

Unlike hypertension, other conditions such as hypothyroidism, hyperlipidaemia, and kidney disease have direct biochemical biomarkers - TSH, comprehensive lipid panels, and urine albumin, respectively - making CGM data largely redundant for their

classification. Retinopathy, nephropathy, and neuropathy, while diabetes complications, may develop through different pathways less directly linked to short-term glycemic variability patterns captured by CGM (Control and Group, 1993).

7.1.2 Opportunistic Screening and Study Objectives

Current hypertension screening relies on periodic blood pressure measurements during clinic visits, which may miss patients with early or borderline hypertension who could benefit from earlier intervention. A key advantage of CGM-based risk classification is that it represents *opportunistic screening*: extracting additional clinical value from data already being collected for glucose management. Unlike approaches requiring additional devices or patient burden, a rich array of CGM data is passively generated as patients manage their diabetes normally, and will only become more important as standard of care shifts toward remote monitoring.

In this study, we analysed CGM data from 689 T1D patients from the T1Diabetes-Granada cohort (Rodriguez-Leon et al., 2023) to test whether adding CGM-derived features to standard biochemical panels improved classification of seven diagnosed comorbidities: hypertension, hypothyroidism, retinopathy, lipid metabolism disorders, airway disease, nephropathy, and neuropathy. We further conducted an ablation study, subgroup analysis, false negative case study, and decision curve analysis to assess the robustness and clinical utility of our findings.

7.2 Methods

7.2.1 Data Source and Study Population

This retrospective cohort study used the T1DiabetesGranada dataset (Rodriguez-Leon et al., 2023), a longitudinal multi-modal dataset of T1D. Participants were patients at the Clinical Unit of Endocrinology and Nutrition of the San Cecilio University

Hospital of Granada, Spain. The study period ran from January 6, 2018 to March 21, 2022, involving 736 patients. CGM was performed using FreeStyle Libre 1 or 2 flash glucose monitoring (FGM) sensors (Abbott Diabetes Care, Inc., Alameda, CA, USA), which capture interstitial glucose readings at 15-minute intervals over a 14-day service life.

The initial cohort consisted of 373 female (50.68%) and 363 male (49.32%) patients, with a mean age of 40.34 ± 15.77 years at the start of data collection (Rodriguez-Leon et al., 2023). The dataset contains over 22.6 million glucose records, with patients contributing an average of 350.24 ± 284.15 monitoring days and 30,802.95 ± 25,704.87 individual measurements (Rodriguez-Leon et al., 2023). The dataset includes four data tables: patient demographics, longitudinal CGM time-series, biochemical laboratory results, and ICD-9 diagnostic codes. The study was approved by the Ethics Committee of Biomedical Research of the Province of Granada (Protocol code: K134665CRL; Ethics portal code: 0698-N-21).

After removing chlorine and number of diagnoses due to high missingness (42.53% and 30.57%, respectively), and excluding patients with missing data, 689 of the 736 original patients remained in the final analysis.

7.2.2 Outcome Definitions

We examined seven comorbidities based on ICD-9 diagnostic codes: (1) hypertension (codes 401, 401.1, 401.9), (2) hypothyroidism (codes 244, 244.3, 244.8, 244.9), (3) retinopathy (codes 250.5, 250.51, 250.53), (4) lipid metabolism disorders (codes 272, 272.1, 272.2, 272.4, 272.8), (5) airway disease (codes 477, 477.9, 493, 493.01, 493.9, 493.91), (6) nephropathy (codes 250.4, 250.41), and (7) neuropathy (codes 250.6, 250.63). Patients were classified as having each comorbidity (binary outcome: 1 = present, 0 = absent) based on the presence of any corresponding diagnostic code in their medical record.

7.2.3 Feature Engineering

CGM-Derived Features

From raw CGM time-series data, we engineered 36 features capturing glycemic levels, variability, and temporal patterns (Rodbard, 2016), where G_i denotes the i-th glucose measurement.

Glycemic variability metrics (9 features): Coefficient of variation (CV) calculated separately for four temporal quartiles (Q1–Q4) of each patient's monitoring period, as well as daytime (6 AM–10 PM) and night-time (10 PM–6 AM) periods. Additionally, Mean Amplitude of Glycemic Excursions (MAGE), Mean of Daily Differences (MODD), and Continuous Overall Net Glycemic Action at 4 time steps apart (CONGA-4) were computed (Service et al., 1970).

$$\mathrm{CV} = \frac{\sigma}{\mu} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(G_i - \mu)^2}}{\mu} \tag{7.1}$$

$$\mathrm{MODD} = \frac{1}{N-1}\sum_{i=1}^{N-1}|G_{i+1} - G_i| \tag{7.2}$$

$$\mathrm{MAGE} = \frac{1}{|\{i : |G_i - G_{i-1}| > \sigma\}|}\sum_{i\,:\,|G_i - G_{i-1}|>\sigma}|G_i - G_{i-1}| \tag{7.3}$$

$$\mathrm{CONGA\text{-}4} = \mathrm{SD}\big(\{G_i - G_{i-4}\}_{i=5}^{N}\big) \tag{7.4}$$

Time-in-range metrics (12 features): Time in range (TIR, 70–180 mg/dL) and time below range (TBR, <70 mg/dL), each calculated separately for four temporal quartiles (Q1–Q4) as well as daytime and night-time periods.

$$\mathrm{TIR} = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[70 \le G_i \le 180]\right) \times 100 \tag{7.5}$$

$$\text{TBR} = \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[G_i < 70] \right) \times 100 \tag{7.6}$$

where $\mathbf{1}[\cdot]$ is the indicator function.

Temporal features (5 features): Linear regression slopes quantifying temporal changes in TIR and mean glucose over the full monitoring period. Standard deviations of rolling means and rolling CVs, capturing medium-term variability patterns across a window of $1/10^{\text{th}}$ the total number of measurements. A composite interaction feature - the Combined Trend-Stability Index - was defined as:

$$\beta_{\text{mean}} = \left(\frac{\sum_{i=1}^{N}(t_i - \bar{t})(G_i - \bar{G})}{\sum_{i=1}^{N}(t_i - \bar{t})^2} \right) \times 86{,}400 \tag{7.7}$$

$$I = \beta_{\text{mean}} \times \overline{M_{\text{quartile}}} \tag{7.8}$$

where β_{mean} is the daily glucose trend slope (mg/dL per day), t_i is the timestamp of measurement G_i, and $\overline{M_{\text{quartile}}}$ is the mean of the quartile-based glycaemic metrics (TIR, TBR, and CV).

Hypoglycemic and hyperglycemic event features (5 features): Counts and average durations of hypoglycemic (<70 mg/dL) and hyperglycemic (>180 mg/dL) episodes, plus a count of rapid glucose changes (>50 mg/dL within consecutive readings).

Distributional percentiles (5 features): Glucose values at the 10th, 25th, 50th, 75th, and 90th percentiles of each patient's distribution.

Variance inflation factor analysis revealed high collinearity among related TIR, TBR, and distributional features. CGM-derived metrics are inherently correlated due to shared physiological structure. XGBoost, a tree-based ensemble method that is robust to linear multicollinearity (Chen and Guestrin, 2016), was therefore used, with model interpretation relying on SHAP-based feature importance.

Biochemical and Demographic Features

We extracted mean values of 15 biochemical parameters: alanine transaminase (ALT), albumin (urine), creatinine (serum and urine), gamma-glutamyl transferase (GGT), glucose, HDL cholesterol, potassium, sodium, thyrotropin (TSH), total cholesterol, triglycerides, uric acid, and HbA1c (mean and standard deviation). Demographic features included age, sex, number of days with CGM measurements, total number of CGM measurements, and number of available biochemical parameters.

7.2.4 Machine Learning Model

We employed XGBoost (Chen and Guestrin, 2016), a gradient boosting algorithm well-suited for tabular medical data. Hyperparameters were selected via grid search with 5-fold stratified cross-validation (ROC-AUC as the optimisation metric), tuning on a pooled subset of patients with retinopathy, airway disease, and nephropathy. Selected hyperparameters were: `n_estimators` = 500, `max_depth` = 2, `learning_rate` = 0.05, `subsample` = 0.8, `colsample_bytree` = 1.0. At each boosting iteration t, XGBoost minimises:

$$L^{(t)} = \sum_{i=1}^{n} l\Big(y_i,\, \hat{y}_i^{(t-1)} + f_t(x_i)\Big) + \Omega(f_t) \tag{7.9}$$

where l is log-loss, $\hat{y}_i^{(t-1)}$ is the current predicted probability from all previous trees, $y_i \in \{0, 1\}$ is the binary label, Ω is a regularisation term penalising model complexity, and f_t is the new tree being fitted.

For each comorbidity, two models were trained: (1) a *biochemical-and-demographic* model using laboratory results and demographics only, and (2) a *combined* model adding all 36 CGM features. Both models used 10-fold stratified cross-validation across 5 random seeds. Within each training fold, the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) was applied to address class imbalance,

while test sets remained unmodified to reflect true population prevalence. SMOTE generates synthetic minority-class samples via:

$$x_{\text{new}} = x + \lambda \times (x_z - x), \quad \lambda \sim \text{Uniform}[0, 1] \tag{7.10}$$

where x is a minority-class sample and x_z is one of its k-nearest neighbours.

7.2.5 Statistical Analysis

Model performance was evaluated using the area under the receiver operating characteristic curve (ROC-AUC). Statistical significance was assessed using DeLong's test ($\alpha = 0.05$), incorporating results from every fold and every seed, testing whether the combined model performed differently from the biochemical-and-demographic model. All analyses were conducted in Python 3.10 using scikit-learn 1.3.0, XGBoost 2.0.3, and imbalanced-learn 0.11.0.

SHAP Analysis

To understand which features drive hypertension risk classification, we employed SHapley Additive exPlanations (SHAP) (Ponce-Bobadilla et al., 2024). SHAP assigns each feature an importance value by computing the weighted average of its marginal contributions across all possible feature subsets:

$$\phi_i(\hat{p}, x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left[\hat{p}(S \cup \{i\}) - \hat{p}(S)\right] \tag{7.11}$$

where ϕ_i is the SHAP value (contribution) of feature i; N is the full feature set (36 CGM metrics + 16 biochemical markers + 6 demographic variables); S is a subset not containing feature i; and $\hat{p}(X)$ is the model's classification probability using features in set X. Global feature importance was aggregated across all 10 folds of seed 42, and a false negative case study of persistently misclassified patients was conducted.

Ablation Study

A feature-group ablation study was performed to quantify the contribution of each CGM feature category to model performance. Using the combined model as baseline, one CGM feature group was removed at a time while retaining all remaining features. The five groups were: glycaemic variability, time-in-range metrics, temporal trends, event-based features, and distributional percentiles. Each ablated model was retrained using the same 10-fold stratified cross-validation, SMOTE procedure, and five random seeds as the primary analysis.

Subgroup Analysis

A subgroup analysis (Lipkovich and Dmitrienko, 2014) was performed by stratifying on age (<45 vs. ≥ 45 years), sex (female vs. male), and HbA1c control ($<7.5\%$ vs. $\geq 7.5\%$). At least 10 positive cases of diagnosed hypertension were required within each stratum to perform the analysis.

7.3 Results

7.3.1 Cohort Characteristics

The final cohort comprised 689 T1D patients. Summary characteristics are presented in Table 7.1. Hypertension was the primary outcome of interest, present in 67 patients (9.7%). Other comorbidities ranged from 2.5% prevalence (neuropathy) to 13.5% (retinopathy).

Table 7.1: Cohort characteristics ($N = 689$).

Characteristic	Value
Demographics	
Age, years	41.34 ± 15.7
Male sex, n (%)	335 (48.6%)
Glycaemic Control	
HbA1c, %	7.89 ± 0.74
CGM Monitoring	
Days with CGM measures, median [IQR]	288 [123–522]
Number of CGM measurements, median [IQR]	25,182 [10,682–45,199]
Clinical Data	
Number of biochemical measurements	124.7 ± 87.8
Comorbidities, n (%)	
Hypertension	67 (9.7%)
Hypothyroidism	91 (13.2%)
Retinopathy	93 (13.5%)
Lipid metabolism disorders	52 (7.5%)
Airway disease	37 (5.4%)
Nephropathy	21 (3.0%)
Neuropathy	17 (2.5%)

Values are mean ± standard deviation, median [IQR], or n (%).

7.3.2 Primary Outcome: Classification of Diagnosed Hypertension

Addition of CGM features to the biochemical-and-demographic model significantly improved classification of diagnosed hypertension (Table 7.2). The biochemical-and-demographic model achieved a mean ROC-AUC of 0.748 ± 0.092 (Fig 7.1). The combined model achieved 0.770 ± 0.088, representing an improvement of ΔAUC = +0.022 ($P = 0.0191$, DeLong's test). A corresponding improvement was observed in precision-recall AUC (PR-AUC: 0.286 ± 0.126 vs. 0.308 ± 0.126).

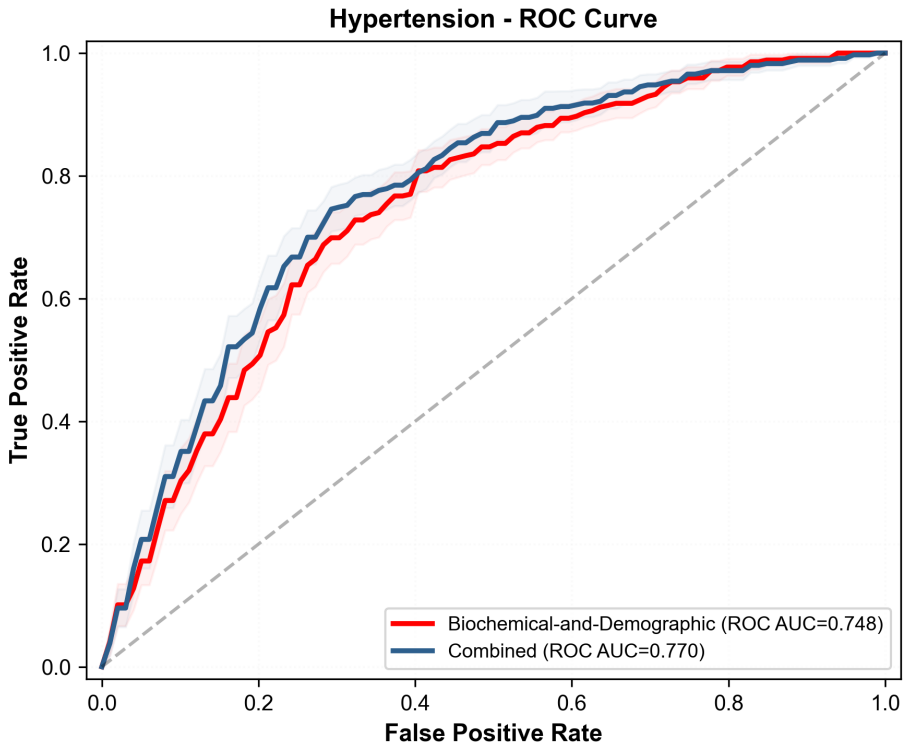

Figure 7.1: Receiver operating characteristic (ROC) curve for hypertension prediction.

7.3.3 SHAP Feature Importance

In the biochemical-and-demographic model, age was the most stable and influential feature (mean $|\text{SHAP}| = 1.22 \pm 0.17$), followed by number of biochemical parameters (0.90 ± 0.10) and albumin (urine) (0.72 ± 0.12). In the combined model, age remained dominant (1.21 ± 0.15), with biochemical features retaining their influence alongside CGM-derived variability metrics. When CGM features were added, they occupied four of the top 15 discriminators: `rolling_cv_std` (0.26 ± 0.09), `rolling_mean_std` (0.21 ± 0.06), `CV_Q4`, and `TBR_Q2`. This suggests that intra-individual glucose variability—rather than absolute glucose levels—contributes meaningfully to hypertension discrimination beyond traditional biochemical markers.

SHAP directionality analysis showed that high values of rolling glucose variability metrics, CV in the fourth quartile, and time below range in the second quartile were associated with increased predicted hypertension risk. Age and renal markers (albumin,

Table 7.2: Model performance for comorbidity classification. Performance metrics are from 10-fold stratified cross-validation across 5 seeds, reported as mean ± SD. *Significant at $\alpha = 0.05$ (DeLong's test).

Comorbidity	Biochem ROC-AUC	Combined ROC-AUC	*P*-value	Biochem PR-AUC	Combined PR-AUC
Hypertension	0.748 ± 0.092	0.770 ± 0.088	0.0191*	0.286 ± 0.126	0.308 ± 0.126
Hypothyroidism	0.697 ± 0.078	0.690 ± 0.082	0.5013	0.310 ± 0.101	0.315 ± 0.096
Retinopathy	0.683 ± 0.072	0.684 ± 0.079	0.8295	0.269 ± 0.078	0.281 ± 0.084
Lipid metabolism	0.738 ± 0.105	0.721 ± 0.107	0.0947	0.267 ± 0.132	0.256 ± 0.134
Airway disease	0.544 ± 0.131	0.534 ± 0.148	0.7429	0.088 ± 0.035	0.123 ± 0.082
Nephropathy	0.851 ± 0.137	0.849 ± 0.126	0.9269	0.470 ± 0.305	0.424 ± 0.277
Neuropathy	0.887 ± 0.088	0.862 ± 0.095	0.0058*	0.306 ± 0.257	0.249 ± 0.242

creatinine) exhibited positive associations with risk at high values, consistent with established clinical relationships.

False Negative Case Study

To investigate sources of model underestimation, we identified five patients with hypertension who were persistently misclassified as false negatives across all five seeds at the 22% decision threshold. Their mean predicted risks ranged from 0.10% to 1.06%, indicating confidently low predicted risk despite true hypertension status.

Four of the five patients were female, with sex exerting a protective SHAP effect in all four cases (`Sex_M` SHAP range: -0.099 to -0.353). Young age below the cohort median was the dominant protective feature in three patients (SHAP range: -0.929 to -1.365), while low urinary albumin dominated in two others (SHAP range: -1.172 to -1.423). CGM-derived features also contributed protective effects across patients, including low rolling glucose variability, favorable time-below-range quartile metrics, and reduced glucose slope. The mean protective SHAP value across all five patients was -0.154, compared to a mean risk-increasing SHAP of $+0.127$.

Counterfactual analysis indicated that replacing each patient's single most protective feature with the population median would increase predicted risk by 0.934 to 1.423 log-odds units, corresponding to approximately a 2.5- to 4-fold increase. These cases highlight systematic model underestimation in younger female patients with

favorable glycemic and metabolic profiles, and point to opportunities for improved calibration in this subgroup.

7.3.4 Ablation Study

Feature-group ablation results are presented in Table 7.3. The combined model was robust to the removal of most individual CGM feature groups, with glycaemic variability, time-in-range, event-based, and distributional groups each producing negligible, non-significant performance changes. Removing the temporal feature group produced the largest performance drop (ΔAUC $= -0.017$, $P < 0.001$), suggesting that longitudinal glucose trend features are the primary CGM contributors to hypertension discrimination.

Table 7.3: Ablation analysis of CGM feature groups. Performance reported for the combined model with the indicated group removed, compared to the full combined model (ROC-AUC $= 0.770 \pm 0.088$). *Significant at $\alpha = 0.05$ (DeLong's test).

Feature Group Removed	Mean ROC-AUC ± SD	ΔAUC	P-value	Sig.
Glycaemic variability (CV, MAGE, MODD, CONGA-4)	0.770 ± 0.085	+0.000	0.7137	
Time in range (TIR, TBR)	0.769 ± 0.090	−0.000	0.9301	
Temporal (slopes, rolling mean, rolling SD)	0.753 ± 0.092	−0.017	<0.001	*
Events (hypo- and hyperglycaemia counts/durations)	0.766 ± 0.092	−0.004	0.3786	
Distribution (10th–90th percentiles)	0.765 ± 0.092	−0.005	0.2268	

7.3.5 Subgroup Analysis

Subgroup analyses across age, sex, and glycemic control were largely null, with no consistent pattern of CGM benefit (Table 7.4). In patients aged ≥45 years, CGM features marginally reduced performance relative to the biochemical-and-demographic model (ΔAUC $= -0.035$, $P = 0.002$), while the <45 group was underpowered ($n_{\text{pos}} = 10$). Sex-stratified results showed negligible differences (ΔAUC $= +0.010$ in males, -0.027 in females), neither reaching significance. Glycemic control stratification produced small deltas (ΔAUC $= -0.021$ for HbA1c <7.5%, $P = 0.013$; ΔAUC $=$

+0.016 for HbA1c ≥7.5%, $P = 0.166$). Overall, no stratum showed robust, statistically significant improvement from CGM addition.

Table 7.4: Subgroup analysis. AUC values reported as mean ± SD. (*) indicates $P < 0.05$.

Variable	Subgroup	N	N_{pos}	Prev. (%)	$AUC_{Biochem}$	$AUC_{Combined}$	ΔAUC	P-value
Age	<45	397	10	2.5	0.610 ± 0.319	0.614 ± 0.364	+0.004	0.9045
Age	≥45	292	57	19.5	0.662 ± 0.103	0.627 ± 0.124	−0.035	0.002(*)
Sex	Female	354	29	8.2	0.687 ± 0.162	0.660 ± 0.160	−0.027	0.1290
Sex	Male	335	38	11.3	0.782 ± 0.085	0.792 ± 0.084	+0.010	0.3331
HbA1c	<7.5%	288	25	8.7	0.838 ± 0.099	0.817 ± 0.120	−0.021	0.0130(*)
HbA1c	≥7.5%	401	42	10.5	0.734 ± 0.107	0.750 ± 0.119	+0.016	0.1663

7.3.6 Decision Curve Analysis

Decision curve analysis demonstrated that the combined model provided greater net benefit than the biochemical-and-demographic model across threshold probabilities of 0.05-0.35 (Figure 7.2). Within this range, the combined model also outperformed the treat-all strategy, and from approximately 0.05 to 0.23, the treat-none strategy as well, indicating clinical utility at thresholds appropriate for a population with 9.7% hypertension prevalence. At a threshold of approximately 0.22, the combined model's net benefit remained positive (net benefit over treat-none = 0.0107, corresponding to approximately 1 additional true positive per 100 patients), whereas the biochemical-and-demographic model's net benefit had dropped below zero (Vickers and Elkin, 2006).

7.3.7 Secondary Outcomes: Other Comorbid Conditions

In contrast to hypertension, CGM data provided no significant improvement for classifying any other comorbid condition (Table 7.2). Differences in ROC-AUC for hypothyroidism, retinopathy, lipid metabolism disorders, airway disease, and nephropathy were all non-significant. For neuropathy, the combined model performed significantly

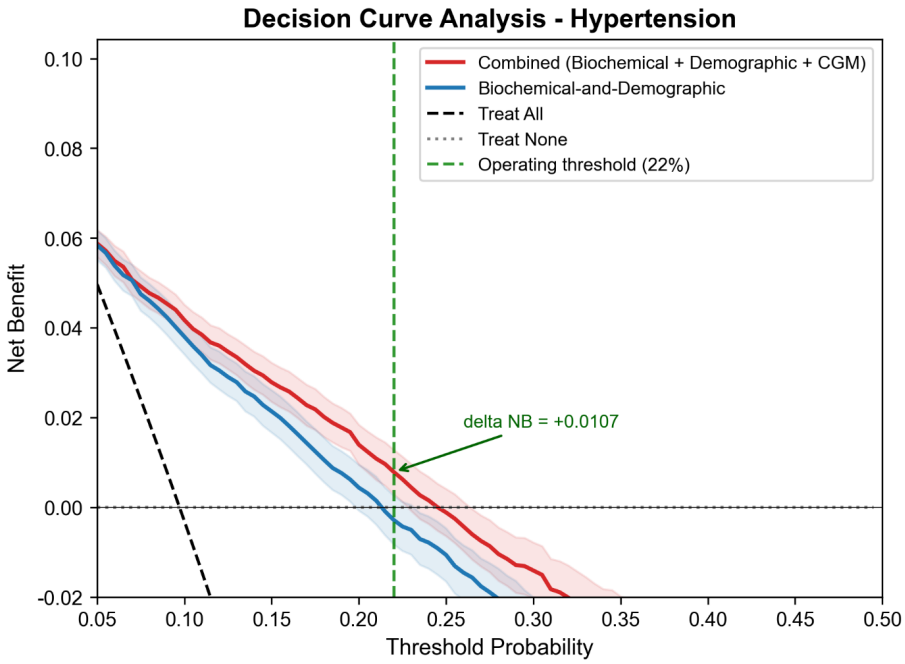

Figure 7.2: Decision curve analysis for the predictive model.

worse than the biochemical-and-demographic model (ΔAUC $= -0.025$, $P = 0.006$), suggesting that CGM features may have added noise for this outcome. Airway disease models performed near random chance (AUC ≈ 0.54), while nephropathy and neuropathy models achieved the highest biochemical-model baseline AUCs (0.851 and 0.887, respectively), reflecting the utility of established biomarkers such as albumin and creatinine for renal and neural complications.

7.4 Discussion

7.4.1 Main Findings

This study demonstrates that CGM data significantly enhances diagnosed hypertension classification in T1D patients when added to standard biochemical and demographic information (ΔAUC $= +0.022$, $P = 0.0191$). Critically, this improvement was specific to hypertension; no significant improvements were observed for hypothyroidism,

retinopathy, lipid disorders, nephropathy, or airway disease, and CGM addition was associated with significantly reduced performance for neuropathy.

Four CGM metrics emerged as top-15 discriminators in the combined model: rolling mean standard deviation, rolling CV standard deviation, CV during the fourth monitoring quartile (`CV_Q4`), and time below range during the second quartile (`TBR_Q2`). The prominence of temporal variability metrics suggests that dynamic glycemic patterns, rather than static glycemic control alone, are meaningfully associated with vascular outcomes. The importance of `TBR_Q2` and `CV_Q4` underscores the potential role of hypoglycemic exposure and late-period glycemic fluctuations, which may reflect the complex interplay between insulin management, physical activity, and autonomic regulation that also influences blood pressure.

7.4.2 Comparison with Prior Literature

Previous studies have shown that CGM data can predict glucose trends (Liu et al., 2025) and that glycemic variability correlates with cardiovascular outcomes (Ceriello et al., 2008). Consistent with prior work linking CGM metrics (including TIR and CV) to all-cause mortality (Okuno et al., 2025), TIR, TBR, CONGA-4, MODD, and CV were all influential discriminators in the combined model. Our findings extend this literature by demonstrating condition-specific discrimination value and quantifying the benefit of CGM data for hypertension specifically. The modest but statistically significant AUC improvement (+0.022) is clinically meaningful given that it requires no additional data collection and could enable earlier identification of at-risk patients (Pencina et al., 2008).

Previous research has established that biochemical and demographic panels can classify comorbid conditions in T1D patients. The Risk Equations for Complications Of type 1 Diabetes (RECODe) study developed and validated comprehensive classification models for retinopathy, albuminuria, and cardiovascular disease using

traditional clinical variables including HbA1c, lipid profiles, blood pressure, diabetes duration, and BMI (Basu et al., 2017). Subsequent studies have confirmed associations between specific biomarkers and individual complications: elevated HbA1c and lipid dysregulation classify diabetic retinopathy and nephropathy (Lu et al., 2018), while thyroid function markers identify patients at risk for hypothyroidism (Kadiyala et al., 2010). The Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications (DCCT/EDIC) cohort demonstrated that long-term glycemic control assessed by HbA1c strongly classifies microvascular complications including neuropathy (Control et al., 2016). However, these traditional models rely on infrequent HbA1c measurements that capture average glucose but may miss glycemic variability - a limitation that CGM-derived metrics may address by providing continuous assessment of glucose fluctuations and time spent in therapeutic range. T1D patients already wear CGM devices for glucose management, generating data that can be analysed for secondary clinical insights.

7.4.3 Clinical Implications

These findings support the development of clinical decision support tools that integrate CGM data for hypertension risk classification. Such tools could be embedded in diabetes management platforms and electronic health records, automatically generating risk scores and flagging patients warranting closer blood pressure monitoring (Bates et al., 2003). This is particularly valuable for young adults with T1D who may have emerging hypertension risk but infrequent monitoring, and for patients in under-resourced settings where access to frequent blood pressure checks is limited (Peiris et al., 2014). For patients identified as high-risk, a structured clinical pathway could include: (1) confirmatory blood pressure measurement at the next clinic visit or via home monitoring; (2) assessment for modifiable risk factors including dietary sodium intake, physical activity, and body weight; (3) more frequent blood pressure monitoring

for those with borderline or high-normal readings; and (4) consideration of earlier antihypertensive therapy initiation in patients meeting diagnostic criteria, particularly those with additional cardiovascular risk factors.

The false negative analysis demonstrates that the model may systematically underestimate hypertension risk in metabolically well-controlled patients whose protective CGM-derived features and favorable lipid profiles dominate the classification. The consistent protective effect of female sex across four of five false negatives warrants further investigation into potential sex-specific miscalibration. These findings underscore the model's capacity to encode nuanced physiological trade-offs - recognising that exceptional glycemic stability and favorable lipid control may partially mitigate age-driven and renal-driven hypertension risk - while highlighting opportunities for improved calibration in younger female patients.

Decision curve analysis contextualises the AUC improvement in terms of clinical decision-making utility. At thresholds exceeding 0.22, the biochemical-and-demographic model yielded negative net benefit, while the combined model remained positive until approximately 0.24, reflecting the established principle that overly conservative screening thresholds in low-prevalence populations can cause net harm by missing true cases at a rate that outweighs the cost of false positives (Vickers and Elkin, 2006). Across a diabetes clinic screening thousands of patients annually, approximately 1 additional true positive per 100 patients screened translates to a meaningful reduction in missed diagnoses at no additional data-collection cost.

7.4.4 Limitations

Temporal data on hypertension onset and medications consumed were unavailable, precluding analysis of incident versus prevalent disease and precluding adjustment for antihypertensive therapy that may have normalised blood pressure values. The diagnosed conditions were recorded at some point during the study visits, but no

information was available on the specific timing of each diagnosis relative to duration of diabetes.

Methodologically, the use of SMOTE to address class imbalance may introduce synthetic patterns not present in real data. The relatively low prevalence of some conditions (e.g., 2.5% for neuropathy) limited statistical power for those analyses. XGBoost, while demonstrating strong discriminative ability on tabular data (Chen and Guestrin, 2016), may be superseded by transformer-based architectures designed for small tabular datasets in the future (Hollmann et al., 2022). Furthermore, the Spanish T1DiabetesGranada cohort may not be fully representative of T1D populations in other healthcare systems, and external validation in independent cohorts is necessary before clinical implementation.

7.4.5 Future Directions

Prospective studies incorporating longitudinal blood pressure measurements could clarify whether CGM patterns predict incident hypertension or blood pressure trajectories over time. Development and clinical trial testing of CGM-based risk scores integrated into clinical workflows would evaluate real-world effectiveness for patients with both T1D and Type 2 diabetes (T2D). Mechanistic studies examining relationships between specific glycemic patterns - such as nocturnal variability and hypoglycaemia frequency - and blood pressure measurements could elucidate underlying pathophysiological pathways. Finally, exploring whether CGM data can classify other cardiovascular outcomes, such as coronary artery disease or stroke, would expand the opportunistic screening paradigm.

7.5 Conclusion

This study demonstrates that CGM data significantly improves classification of diagnosed hypertension in T1D patients when added to standard biochemical and demographic information, with no additional patient burden. The specificity of this improvement to hypertension, but not to other comorbidities, suggests distinct pathophysiological links between glycemic variability patterns and blood pressure regulation. This represents an opportunistic screening approach that leverages existing monitoring infrastructure to extract additional clinical value from the vast array of CGM data continuously generated during routine diabetes management. As wearable health technologies become increasingly prevalent, these findings support a paradigm shift toward multi-purpose use of device data for comprehensive health monitoring and early disease detection in T1D - and potentially T2D - patients.

Chapter 8

Contributions

The contributions of this dissertation are methodological and applied. Chapter 2 introduced a novel MCMC algorithm that combined AR(1) and negative binomial likelihood, applied to COVID-19 daily cases, in a way that has not been done before, and can be extended to multiple changepoints, additional autoregressive values, and different application areas. Chapter 3 introduced a penalized fractional polynomial model that can be fit with relative ease using standard mixed model tools in R. Chapter 4 showed that XGBoost was the best performing model, along with Random Forest, in classifying hospital readmission for patients with diabetes. Chapter 5 created a novel stacking ensemble approach using XGBoost as the meta-learner, achieving better predictive performance than the base learners. Chapter 6 creates a novel generative model that can be used to create treatment protocols for patients with sepsis, and in fact any disease with longitudinal treatment, in the absence of stepwise rewards. Chapter 7 paves the way for repurposing CGM into a tool for hypertension screening, leading to comprehensive healthcare with no additional patient burden.

Bibliography

Alkanhel, R. I., Saleh, H., Elaraby, A., Alharbi, S., Elmannai, H., Alaklabi, S., Alsamhi, S. H., and Mostafa, S. (2024). Hybrid cnn-gru model for real-time blood glucose forecasting: Enhancing iot-based diabetes management with ai. *Sensors*, 24(23):7670.

Atkeson, C. G. and Schaal, S. (1997). Robot learning from demonstration. In *ICML*, volume 97, pages 12–20.

Barry, D. and Hartigan, J. A. (1993). A bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):309–319.

Basu, S., Sussman, J. B., Berkowitz, S. A., Hayward, R. A., and Yudkin, J. S. (2017). Development and validation of risk equations for complications of type 2 diabetes (recode) using individual participant data from randomised trials. *The lancet Diabetes & endocrinology*, 5(10):788–798.

Bates, D. W., Kuperman, G. J., Wang, S., Gandhi, T., Kittler, A., Volk, L., Spurr, C., Khorasani, R., Tanasijevic, M., and Middleton, B. (2003). Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. *Journal of the American Medical Informatics Association*, 10(6):523–530.

Benbassat, J. and Taragin, M. (2000). Hospital readmissions as a measure of quality of health care: advantages and limitations. *Archives of internal medicine*, 160(8):1074–1081.

Bergman, R. N. (2021). Origins and history of the minimal model of glucose regulation. *Frontiers in endocrinology*, 11:583016.

Bian, Q., As' arry, A., Cong, X., Rezali, K. A. b. M., and Raja Ahmad, R. M. K. b. (2024). A hybrid transformer-lstm model apply to glucose prediction. *PLoS One*, 19(9):e0310084.

Byrne, L. and Van Haren, F. (2017). Fluid resuscitation in human sepsis: time to rewrite history? *Annals of intensive care*, 7(1):4.

Ceriello, A., Esposito, K., Piconi, L., Ihnat, M. A., Thorpe, J. E., Testa, R., Boemi, M., and Giugliano, D. (2008). Oscillating glucose is more deleterious to endothelial function and oxidative stress than mean glucose in normal and type 2 diabetic patients. *Diabetes*, 57(5):1349–1354.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chen, J. and Gupta, A. K. (1997). Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical Association*, 92(438):739–747.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. (2021). Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chepulis, L., Morison, B., Cassim, S., Norman, K., Keenan, R., Paul, R., and Lawrenson, R. (2021). Barriers to diabetes self-management in a subset of new zealand adults with type 2 diabetes and poor glycaemic control. *Journal of diabetes research*, 2021(1):5531146.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86(2):221–241.

Choi, Y., Oh, S., Huh, J. W., Joo, H.-T., Lee, H., You, W., Bae, C.-m., Choi, J.-H., and Kim, K.-J. (2024). Deep reinforcement learning extracts the optimal sepsis treatment policy from treatment records. *Communications medicine*, 4(1):245.

Choudhary, K., Gupta, D., and Thomas, P. S. (2024). Icu-sepsis: A benchmark mdp built from real medical data. *arXiv preprint arXiv:2406.05646*.

Cichosz, S. L. and Schaarup, C. (2017). Hyperglycemia as a predictor for adverse outcome in icu patients with and without diabetes. *Journal of Diabetes Science and Technology*, 11(6):1272–1273.

Committee, A. D. A. P. P. et al. (2023). 6. glycemic goals and hypoglycemia: standards of care in diabetes—2024. *Diabetes care*, 47(Suppl 1):S111.

Control, D. and Group, C. T. R. (1993). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *New England journal of medicine*, 329(14):977–986.

Control, D., of Diabetes Interventions, C. T. D., and Group, C. E. S. R. (2016). Intensive diabetes treatment and cardiovascular outcomes in type 1 diabetes: the dcct/edic study 30-year follow-up. *Diabetes care*, 39(5):686–693.

Cui, S., Wang, D., Wang, Y., Yu, P.-W., and Jin, Y. (2018). An improved support

vector machine-based diabetic readmission prediction. *Computer methods and programs in biomedicine*, 166:123–135.

Danne, T., Nimri, R., Battelino, T., Bergenstal, R. M., Close, K. L., DeVries, J. H., Garg, S., Heinemann, L., Hirsch, I., Amiel, S. A., et al. (2017). International consensus on use of continuous glucose monitoring. *Diabetes care*, 40(12):1631–1640.

Dantes, R. B. (2023). Sepsis program activities in acute care hospitals—national healthcare safety network, united states, 2022. *MMWR. Morbidity and mortality weekly report*, 72.

Davis, R. A. and Liu, H. (2012). Theory and inference for a class of observation-driven models with application to time series of counts. *arXiv preprint arXiv:1204.3915*.

De Ferranti, S. D., De Boer, I. H., Fonseca, V., Fox, C. S., Golden, S. H., Lavie, C. J., Magge, S. N., Marx, N., McGuire, D. K., Orchard, T. J., et al. (2014). Type 1 diabetes mellitus and cardiovascular disease: a scientific statement from the american heart association and american diabetes association. *Circulation*, 130(13):1110–1130.

Fagherazzi, G. and Ravaud, P. (2019). Digital diabetes: Perspectives for diabetes prevention, management and research. *Diabetes & metabolism*, 45(4):322–329.

Farrington, C. P., Andrews, N. J., Beale, A. D., and Catchpole, M. A. (1996). A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society: Series A*, 159(3):547–563.

Fearnhead, P. (2006). Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16(2):203–213.

Fokianos, K., Rahbek, A., and Tjøstheim, D. (2009). Poisson autoregression. *Journal of the American Statistical Association*, 104(488):1430–1439.

Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. (2020). D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*.

Fu, X., Wang, Y., Cates, R. S., Li, N., Liu, J., Ke, D., Liu, J., Liu, H., and Yan, S. (2023). Implementation of five machine learning methods to predict the 52-week blood glucose level in patients with type 2 diabetes. *Frontiers in endocrinology*, 13:1061507.

Goel, S. and Sharma, S. (2022). A comparative study of time series models for blood glucose prediction. In *Proceedings of the Third International Conference on Information Management and Machine Intelligence: ICIMMI 2021*, pages 81–91. Springer.

Gómez-Castillo, N. Y., Cajilima-Cardenaz, P. E., Zhinin-Vera, L., Maldonado-Cuascota, B., León Domínguez, D., Pineda-Molina, G., Hidalgo-Parra, A. A., and Gonzales-Zubiate, F. A. (2021). A machine learning approach for blood glucose level prediction using a lstm network. In *International Conference on Smart Technologies, Systems and Applications*, pages 99–113. Springer.

Green, A., Hede, S. M., Patterson, C. C., Wild, S. H., Imperatore, G., Roglic, G., and Beran, D. (2021). Type 1 diabetes in 2017: global estimates of incident and prevalent cases in children and adults. *Diabetologia*, 64(12):2741–2750.

Gregory, N. S., Seley, J. J., Dargar, S. K., Galla, N., Gerber, L. M., and Lee, J. I. (2018). Strategies to prevent readmission in high-risk patients with diabetes: the importance of an interdisciplinary approach. *Current diabetes reports*, 18(8):54.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182.

Hai, A. A., Weiner, M. G., Paranjape, A., Livshits, A., Brown, J. R., Obradovic, Z., and Rubin, D. J. (2023). Deep learning vs traditional models for predicting

hospital readmission among patients with diabetes. In *AMIA Annual Symposium Proceedings*, volume 2022, page 512.

Hawkins, D. M. and Olwell, D. H. (1998). *Cumulative Sum Charts and Charting for Quality Improvement*. Springer, New York.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. (2022). Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*.

Hovorka, R., Canonico, V., Chassin, L. J., Haueter, U., Massi-Benedetti, M., Federici, M. O., Pieber, T. R., Schaller, H. C., Schaupp, L., Vering, T., et al. (2004). Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. *Physiological measurement*, 25(4):905–920.

Huang, Y., Yang, Z., and Rahmani, A. (2025). Mimic-sepsis: A curated benchmark for modeling and learning from sepsis trajectories in the icu. In *2025 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–7. IEEE.

Janner, M., Du, Y., Tenenbaum, J. B., and Levine, S. (2022). Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*.

Jiang, H. J., Stryer, D., Friedman, B., and Andrews, R. (2003). Multiple hospitalizations for patients with diabetes. *Diabetes care*, 26(5):1421–1426.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Jung, R. C. and Tremayne, A. (2006). Binomial thinning models for integer time series. *Statistical Modelling*, 6(2):81–96.

Kadiyala, R., Peter, R., and Okosieme, O. E. (2010). Thyroid dysfunction in patients with diabetes: clinical implications and screening strategies. *International journal of clinical practice*, 64(8):1130–1139.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and structural biotechnology journal*, 15:104–116.

Khanna, D., Clements, P. J., Volkmann, E. R., Wilhalme, H., Tseng, C.-h., Furst, D. E., Roth, M. D., Distler, O., and Tashkin, D. P. (2019). Minimal clinically important differences for the modified rodnan skin score: results from the scleroderma lung studies (sls-i and sls-ii). *Arthritis research & therapy*, 21(1):23.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Knoblauch, J., Jewson, J. E., and Damoulas, T. (2018). Doubly robust bayesian inference for non-stationary streaming data with β-divergences. *Advances in Neural Information Processing Systems*, 31.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720.

Kopecky, S. and Dwyer, C. (2023). Nature inspired metaheuristic techniques of firefly

and grey wolf algorithms implemented in phishing intrusion detection systems. In *Science and Information Conference*, pages 1309–1332. Springer.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. (2020). Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33:1179–1191.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

Lamari, M., Boukhamla, A., Lafia, T. A., Azizi, N., Kouadria, L., and Hammami, N. E. (2024). Improved ensemble-based approaches with stacking for imbalanced medical data classification. In *2024 1st International Conference on Electrical, Computer, Telecommunication and Energy Technologies (ECTE-Tech)*, pages 1–8. IEEE.

Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 209–225.

Leimkuhler, B. and Matthews, C. (2013). Rational construction of stochastic numerical methods for molecular sampling. *Applied Mathematics Research eXpress*, 2013(1):34–56.

Lewis, S. M. and Raftery, A. E. (1999). Bayesian analysis of event history models with unobserved heterogeneity via markov chain monte carlo: application to the explanation of fertility decline. *Sociological Methods & Research*, 28(1):35–60.

Lipkovich, I. and Dmitrienko, A. (2014). Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using sides. *Journal of biopharmaceutical statistics*, 24(1):130–153.

Liu, V. B., Sue, L. Y., Padilla, O. M., and Wu, Y. (2025). Optimizing blood glucose predictions in type 1 diabetes patients using a stacking ensemble approach. *Endocrine and Metabolic Science*, 18:100253.

Liu, Y., Yao, S., Shan, X., Luo, Y., Yang, L., Dai, W., and Hu, B. (2024). Time trends and advances in the management of global, regional, and national diabetes in adolescents and young adults aged 10–24 years, 1990–2021: analysis for the global burden of disease study 2021. *Diabetology & Metabolic Syndrome*, 16(1):252.

Liventsev, V. and Fritz, T. (2024). Intensive care as one big sequence modeling problem. *arXiv preprint arXiv:2402.17501*.

Lu, J., Ma, X., Zhou, J., Zhang, L., Mo, Y., Ying, L., Lu, W., Zhu, W., Bao, Y., Vigersky, R. A., et al. (2018). Association of time in range, as assessed by continuous glucose monitoring, with diabetic retinopathy in type 2 diabetes. *Diabetes care*, 41(11):2370–2376.

Mayo, M., Chepulis, L., and Paul, R. G. (2019). Glycemic-aware metrics and oversampling techniques for predicting blood glucose levels using machine learning. *PloS one*, 14(12):e0225613.

McKay, M. D., Rome, M. M., and Bernt, F. M. (1997). Disease chronicity and quality of care in hospital readmissions. *The Journal for Healthcare Quality (JHQ)*, 19(2):33–37.

Mirjalili, S., Mirjalili, S. M., and Lewis, A. (2014). Grey wolf optimizer. *Advances in engineering software*, 69:46–61.

Monnier, L., Mas, E., Ginet, C., Michel, F., Villon, L., Cristol, J.-P., and Colette, C. (2006). Activation of oxidative stress by acute glucose fluctuations compared with sustained chronic hyperglycemia in patients with type 2 diabetes. *Jama*, 295(14):1681–1687.

Moss, S. E., Klein, R., and Klein, B. E. (1999). Risk factors for hospitalization in people with diabetes. *Archives of internal medicine*, 159(17):2053–2057.

Nijkamp, E., Hill, M., Zhu, S.-C., and Wu, Y. N. (2019). Learning non-convergent non-persistent short-run mcmc toward energy-based model. *Advances in Neural Information Processing Systems*, 32.

Okuno, T., Macwan, S. A., Norman, G. J., Miller, D. R., Reaven, P. D., and Zhou, J. J. (2025). Continuous glucose monitoring metrics predict all-cause mortality in diabetes: a real-world long-term study. *Diabetes Care*, 48(10):1794–1802.

Parker, E. D., Lin, J., Mahoney, T., Ume, N., Yang, G., Gabbay, R. A., ElSayed, N. A., and Bannuru, R. R. (2024). Economic costs of diabetes in the us in 2022. *Diabetes care*, 47(1):26–43.

Peiris, D., Praveen, D., Johnson, C., and Mogulluru, K. (2014). Use of mhealth systems and tools for non-communicable diseases in low-and middle-income countries: a systematic review. *Journal of cardiovascular translational research*, 7(8):677–691.

Pencina, M. J., D'Agostino Sr, R. B., D'Agostino Jr, R. B., and Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond. *Statistics in medicine*, 27(2):157–172.

Peng, X., Ding, Y., Wihl, D., Gottesman, O., Komorowski, M., Lehman, L.-w. H., Ross, A., Faisal, A., and Doshi-Velez, F. (2018). Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. In *AMIA Annual Symposium Proceedings*, volume 2018, page 887.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Springer.

Pompa, M., Panunzi, S., Borri, A., and De Gaetano, A. (2021). A comparison among three maximal mathematical models of the glucose-insulin system. *PloS one*, 16(9):e0257789.

Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S., and Stodtmann, S. (2024). Practical guide to shap analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and translational science*, 17(11):e70056.

Prioleau, T., Bartolome, A., Comi, R., and Stanger, C. (2023). Diatrend: A dataset from advanced diabetes technology to enable development of novel analytic solutions. *Scientific Data*, 10(1):556.

QingXiang, B., As' array, A., XiangGuo, C., bin Md Rezali, K. A., and bin Raja Ahmad, R. M. K. (2023). The modified arima predicting algorithm apply on glucose values prediction. In *International Conference on Electrical, Control & Computer Engineering*, pages 25–34. Springer.

Raghu, A., Komorowski, M., Celi, L. A., Szolovits, P., and Ghassemi, M. (2017). Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Machine learning for healthcare conference*, pages 147–163. PMLR.

Rajput, G. and Alashetty, A. (2022). A machine learning approach to reduce the diabetes patient's readmission risk using a novel preprocessing technique. In *2022 4th International Conference on Circuits, Control, Communication and Computing (I4C)*, pages 173–177. IEEE.

Reno, C. M., Daphna-Iken, D., Chen, Y. S., VanderWeele, J., Jethi, K., and Fisher, S. J. (2013). Severe hypoglycemia–induced lethal cardiac arrhythmias are mediated by sympathoadrenal activation. *Diabetes*, 62(10):3570–3581.

Riddle, M. C., Blonde, L., Gerstein, H. C., Gregg, E. W., Holman, R. R., Lachin, J. M., Nichols, G. A., Turchin, A., and Cefalu, W. T. (2019). Diabetes care editors' expert forum 2018: managing big data for diabetes research and care.

Rodbard, D. (2016). Continuous glucose monitoring: a review of successes, challenges, and opportunities. *Diabetes technology & therapeutics*, 18(2_suppl):S2–3.

Rodriguez-Leon, C., Aviles-Perez, M. D., Banos, O., Quesada-Charneco, M., Lopez-Ibarra Lozano, P. J., Villalonga, C., and Munoz-Torres, M. (2023). T1diabetes-granada: a longitudinal multi-modal dataset of type 1 diabetes mellitus. *Scientific Data*, 10(1):916.

Rodríguez-Rodríguez, I., Campo-Valera, M., Rodríguez, J.-V., and Woo, W. L. (2024). Iomt innovations in diabetes management: Predictive models using wearable data. *Expert Systems with Applications*, 238:121994.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Ross, S., Gordon, G., and Bagnell, D. (2011). A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings.

Royston, P. (2017). Model selection for univariable fractional polynomials. *The Stata Journal*, 17(3):619–629.

Royston, P. and Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 43(3):429–453.

Royston, P., Ambler, G., and Sauerbrei, W. (1999). The use of fractional polynomials to model continuous risk variables in epidemiology. *International journal of epidemiology*, 28(5):964–974.

Rubin, D. J. (2018). Correction to: hospital readmission of patients with diabetes. *Current diabetes reports*, 18(4):21.

Ruggieri, E. (2013). A bayesian approach to detecting change points in climatic records. *International Journal of Climatology*, 33(2):520–528.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Number 12. Cambridge university press.

Sauerbrei, W., Royston, P., and Look, M. (2007). A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. *Biometrical Journal*, 49(3):453–473.

Sendak, M. P., Ratliff, W., Sarro, D., Alderton, E., Futoma, J., Gao, M., Nichols, M., Revoir, M., Yashar, F., Miller, C., et al. (2020). Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. *JMIR medical informatics*, 8(7):e15182.

Service, F. J., Molnar, G. D., Rosevear, J. W., Ackerman, E., Gatewood, L. C., and Taylor, W. F. (1970). Mean amplitude of glycemic excursions, a measure of diabetic instability. *Diabetes*, 19(9):644–655.

Shang, Y., Jiang, K., Wang, L., Zhang, Z., Zhou, S., Liu, Y., Dong, J., and Wu, H. (2021). The 30-days hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers. *BMC medical informatics and decision making*, 21(Suppl 2):57.

Shaw, J. L., Bannuru, R. R., Beach, L., ElSayed, N. A., Freckmann, G., Füzéry, A. K., Fung, A. W., Gilbert, J., Huang, Y., Korpi-Steiner, N., et al. (2024). Consensus considerations and good practice points for use of continuous glucose monitoring systems in hospital settings. *Diabetes Care*, 47(12):2062–2075.

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J.-D., Coopersmith, C. M., et al. (2016). The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810.

Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., and Clore, J. N. (2014). Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014(1):781670.

Stratton, I. M., Adler, A. I., Neil, H. A. W., Matthews, D. R., Manley, S. E., Cull, C. A., Hadden, D., Turner, R. C., and Holman, R. R. (2000). Association of glycaemia with macrovascular and microvascular complications of type 2 diabetes (ukpds 35): prospective observational study. *bmj*, 321(7258):405–412.

Swain, A., Ganatra, V., Saha, S., Mathur, A., and Phadke, R. (2022). P-lstm: a novel lstm architecture for glucose level prediction problem. In *International Conference on Neural Information Processing*, pages 369–380. Springer.

Thenappan, S., Ramshankar, N., Hemavathy, N., Subashree, V., Manjul, R. R., and Rajeswari, J. (2023). Machine learning classifiers to decrease diabetic patients probability of hospital readmission. In *2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 829–834. IEEE.

Thorn, L. M., Forsblom, C., Fagerudd, J., Thomas, M. C., Pettersson-Fernholm, K.,

Saraheimo, M., WADen, J., Ronnback, M., Rosengård-Bärlund, M., Björkesten, C.-G. a., et al. (2005). Metabolic syndrome in type 1 diabetes: association with diabetic nephropathy and glycemic control (the finndiane study). *Diabetes care*, 28(8):2019–2024.

Vickers, A. J. and Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574.

Wang, D. et al. (2021). A comparison of machine learning methods to predict hospital readmission of diabetic patient. *Studies of Applied Economics*, 39(4).

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. chapman and hall/CRC.

Yamagata, T., Khalil, A., and Santos-Rodriguez, R. (2023). Q-learning decision transformer: Leveraging dynamic programming for conditional sequence modelling in offline rl. In *International Conference on Machine Learning*, pages 38989–39007. PMLR.

Yang, H., Chen, Z., Huang, J., and Li, S. (2024). Awd-stacking: An enhanced ensemble learning model for predicting glucose levels. *Plos one*, 19(2):e0291594.

Yang, J., Li, L., Shi, Y., and Xie, X. (2018). An arima model with adaptive orders for predicting blood glucose concentrations and hypoglycemia. *IEEE journal of biomedical and health informatics*, 23(3):1251–1260.

Ying, L. P., Yin, O. X., Quan, O. W., Jain, N., Mayuren, J., Pandey, M., Gorain, B., and Candasamy, M. (2025). Continuous glucose monitoring data for artificial intelligence-based predictive glycemic event: A potential aspect for diabetic care. *International Journal of Diabetes in Developing Countries*, 45(2):272–287.

Zapatero, A., Gómez-Huelgas, R., González, N., Canora, J., Asenjo, Á., Hinojosa, J., Plaza, S., Marco, J., and Barba, R. (2014). Frequency of hypoglycemia and its impact on length of stay, mortality, and short-term readmission in patients with diabetes hospitalized in internal medicine wards. *Endocrine Practice*, 20(9):870–875.

Zeger, S. L. and Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics*, pages 1019–1031.

Zhou, B., Rayner, A. W., Gregg, E. W., Sheffer, K. E., Carrillo-Larco, R. M., Bennett, J. E., Shaw, J. E., Paciorek, C. J., Singleton, R. K., Pires, A. B., et al. (2024). Worldwide trends in diabetes prevalence and treatment from 1990 to 2022: a pooled analysis of 1108 population-representative studies with 141 million participants. *The Lancet*, 404(10467):2077–2093.

Zhu, F. (2012). Modeling overdispersed or underdispersed count data with generalized poisson integer-valued garch models. *Journal of Mathematical Analysis and Applications*, 389(1):58–71.

Zhu, T., Li, K., Kuang, L., Herrero, P., and Georgiou, P. (2020). An insulin bolus advisor for type 1 diabetes using deep reinforcement learning. *Sensors*, 20(18):5058.