Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
Università degli Studi di Padova

Head Office: Università degli Studi di Padova

Department of Cardiac, Thoracic, Vascular Sciences and Public Health

Ph.D. Course in: Translational Specialistic Medicine "G. B. Morgagni"

Curriculum: Biostatistics and Clinical Epidemiology

Series: XXXVIII

Monitoring Patient Reported Outcomes and Neurophysiologic Signals by Integrating Clinical Records with High-Dimensional Digital Data

Thesis written with the financial contribution of the European Union - Next Generation EU and *Zeta Research Srl.*

Coordinator: Professor Dario Gregori
Supervisor: Professor Dario Gregori
Co-Supervisor: Dr. Giulia Lorenzoni

Ph.D. student: Noor Muhammad Khan

Abstract

Modern healthcare generates heterogeneous data streams that span electronic health records (EHR), patient-reported outcomes (PROs), and high-frequency neurophysiologic signals. This thesis develops and applies an integrated statistical framework that combines Frequentist and Bayesian approaches to monitor patient outcomes across these modalities while preserving interpretability and rigorous uncertainty quantification.

Chapter 2 establishes the methodological foundation in an infectious-disease surveillance setting. We analyze linked clinical data using generalized linear models with penalization and Bayesian counterparts for shrinkage and partial pooling; we extend to correlated data via generalized linear mixed models, generalized estimating equations, and Bayesian hierarchical models; we diagnose misspecification and overdispersion for binary and count endpoints; address missingness with multiple imputation and fully Bayesian joint modeling; and we evaluate policy effects using interrupted time series with classical and Bayesian ARIMA, as well as structural time series models. Together, these components illustrate how complementary paradigms support transparent estimation, probabilistic prediction, and principled sensitivity analysis in real-world public-health applications.

Chapter 3 turns to high-dimensional intracranial EEG. We perform physiologically guided feature extraction using *FOOOF* to parameterize each power spectrum into aperiodic and oscillatory components, then rank regional stability via within-subject coefficients of variation using Friedman tests and Kendall's W. Hierarchical Bayesian models (fitted in R/Stan) quantify region- and time-of-day effects while accounting for subject and channel heterogeneity. Posterior predictive checks and PSIS–LOO confirm model adequacy. Results reveal robust stratification of spectral-stability across regions and a positive residual correlation between aperiodic offset and exponent, supporting joint modeling.

Chapter 4 addresses perioperative neurophysiology in pediatric anesthesia using the Patient State Index (PSI). We detect abrupt state changes with a two-tier pipeline—global segmentation via PELT and patient-wise Bayesian structural time series—then summarize instability with a phase-normalized, probability-based *Variability Ratio Index* (VARI). Population-averaged inference uses a logistic GEE with a data-driven working correlation and robust standard errors. Instability concentrates around surgical transitions, whereas other static characteristics (e.g., gender or ethnicity) have little effect. In contrast, children of older ages and larger body sizes (each considered independently) showed significantly more stable PSI dynamics.

Across domains, this thesis demonstrates a reproducible, R-first workflow that unifies interpretable feature engineering with hierarchical modeling and probability-focused monitoring. This framework is potentially generalizable to other multi-modal clinical contexts where rigorous uncertainty quantification, biological plausibility, and operational interpretability are essential for data-driven decision-making.

List of Publications

Publications (accepted)

Book

1. **Khan, N. M.**, Baldi, I., Chiaruttini, M. V., & Gregori, D. (in press, 2025). *Classical and Bayesian Statistical Approaches in Infectious Disease Data Analysis*. Springer, Cham. Open Access. https://link.springer.com/book/9783032067463

   Chapters in this volume:

   Chapter 1:. Bayesian and Frequentist Approaches in Infectious Disease Data Analysis.

   Chapter 2:. Generalized Linear Models in Infectious Disease Analysis and Surveillance: Methods for Independent Data .

   Chapter 3:. Variable Selection in Generalized Linear Models.

   Chapter 4:. Machine Learning Models for Probabilistic Inference and Prediction.

   Chapter 5:. Generalized Linear Models in Infectious Disease Analysis and Surveillance: Methods for Correlated Data.

   Chapter 6:. Residuals and Overdispersion in Generalized Linear Models.

   Chapter 7:. Interrupted Time Series Model in Infectious Disease Research and Surveillance.

   Chapter 8:. Generalized Linear Models with Missing Data.

Publications (under processing)

Journal articles

1. Quantifying Regional Variability in Neural Power Spectra: Coefficient of Variation Analysis and Bayesian Multilevel Modeling.

2. A new statistical index for evaluating variability in Patient State Index during pediatric anesthesia.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Modern healthcare data are increasingly multi-modal, combining traditional clinical records with patient-generated information and continuous physiological measurements. Patient-reported outcomes (PROs), such as symptom scores, quality-of-life surveys, and other self-reported measures, provide critical insight into the patient's perspective and are being integrated into electronic health records to enhance patient-centered care [2]. Likewise, neurophysiologic signals captured via high-dimensional digital devices (e.g., wearable or bedside EEG monitors) offer objective, real-time indicators of patient state. Managing these rich data sources in tandem promises a more holistic approach to monitoring health and disease progression. However, the heterogeneity, high dimensionality, and temporal complexity of such data present significant analytical challenges. This doctoral work addresses these challenges by developing and applying statistical methods that integrate clinical records with PROs and neurophysiologic signals to improve the monitoring of patient outcomes. A unifying strategy is the combined use of Frequentist and Bayesian inference paradigms as complementary tools for robust analysis. Frequentist methods provide transparent estimates and confidence intervals with nominal coverage (e.g., 95% under correct model assumptions), while Bayesian models incorporate prior knowledge and hierarchical structures to better handle complex and sparse data. Rather than treating these paradigms as competitors, we employ them as a coherent toolkit [3] that leverages the strengths of each — for example, using Frequentist approaches for initial descriptive analysis and model diagnostics, and Bayesian approaches for multilevel modeling and probabilistic prediction, to take advantage of Bayesian partial pooling and probability statements. This dual perspective is particularly valuable when integrating disparate data types (EHR fields, PRO questionnaires, high-frequency signals), where borrowing information across sources and quantifying uncertainty with explicit probability statements can greatly aid interpretability and decision-making [3]. The following chapters (Two, Three, and Four) each target a different domain within this theme, collectively advancing statistical methods for monitoring patient outcomes and neurophysiology in data-rich environments.

Chapter Two establishes the methodological foundation by comparing and blending classical and Bayesian statistical approaches in the context of applied health data. Titled "Classical and Bayesian Statistics in Infectious Disease Analysis and Surveillance," this chapter illustrates how a broad arsenal of models can be deployed on integrated clinical datasets. It begins by reviewing the fundamentals of classical and Bayesian statistics. It

then covers generalized linear models (GLM) for cross-sectional outcomes, extending to mixed-effects and longitudinal models, i.e., generalized linear mixed models (GLMM) and generalized estimating equations (GEE), to account for clustered and repeated measures. Throughout, the chapter emphasizes the synergy between Frequentist and Bayesian paradigms; for instance, classical estimators (maximum likelihood estimates, likelihood ratio tests) are obtained for transparency and baseline inference, while Bayesian models are constructed in parallel to incorporate prior information and perform hierarchical partial pooling across strata [3]. Techniques for variable selection and regularization in high-dimensional settings are also introduced, including penalized regression methods like the LASSO [4]. This demonstrates how shrinkage priors or penalties can improve prediction without sacrificing interpretability. In addition, Chapter Two outlines strategies for probabilistic prediction using modern machine-learning-influenced tools (naive Bayes classifiers, Bayesian networks, Bayesian additive regression trees, etc.) and discusses the rigorous treatment of missing data through multiple imputations and Bayesian sensitivity analyses. At the end, it also extends to time-series methods (e.g., ARIMA models and their Bayesian counterparts), presenting how they can be used for disease surveillance signals and how policy interventions can be evaluated with interrupted time-series designs. To ground these methods in reality, the chapter presents a case study using data from the "All of Us" Research Program [5]. In this example, diverse data streams (electronic health records, patient surveys, lab results, and more) are linked to study an infectious-disease surveillance problem. Specifically, an individual-level cohort is constructed to examine factors (demographics, comorbidities, prior utilization, etc.) associated with antibiotic prescribing after viral illnesses, while a complementary aggregate dataset tracks monthly antibiotic prescription rates following viral diagnoses. Using this rich dataset, Chapter Two demonstrates how the combined toolkit of GLM/GLMM and Bayesian hierarchical modeling can yield actionable insights—for example, identifying patient characteristics tied to antibiotic use and evaluating the impact of a national antibiotic stewardship intervention through an interrupted time-series analysis [6]. The analysis showcases the benefits of integrating multi-source health data (including PRO-like survey inputs) with rigorous statistical modeling: classical methods ensure that results are reproducible and easily communicated, whereas Bayesian techniques provide deeper uncertainty quantification and accommodate the multi-level structure of the data. In summary, Chapter Two's contribution lies in formulating an integrative statistical framework for health surveillance data, illustrating how careful model selection, evidence synthesis, and validation can translate heterogeneous patient data into meaningful public health conclusions.

Chapter Three, titled "Quantifying Regional Variability in Neural Power Spectra: Stability Mapping and Bayesian Multilevel Modeling," shifts the focus to high-dimensional neurophysiological data by presenting a case study in advanced EEG signal analysis and hierarchical modeling. The central question is how to monitor and characterize neural stability across different brain structures. To tackle the complexity of iEEG (with its multi-level nested structure), Chapter Three develops a workflow that combines domain-guided feature extraction with multi-level statistical modeling [7]. First, after standard data preprocessing and structuring, a parametric spectral decomposition technique known as FOOOF ("Fitting Oscillations and One-Over-F") is applied to each EEG segment [7]. This method separates the power spectrum into a broadband aperiodic component and discrete oscillatory peaks, yielding interpretable features such as the aperiodic baseline

(offset), slope (exponent), and band-specific power with its location (center frequency) of neural oscillations. By reducing each 20-second window of iEEG to a small set of physiologically meaningful parameters, we achieve substantial data reduction without losing interpretability. This is an important advantage over conventional PCA or black-box feature extraction, as the FOOOF features map directly onto known neurophysiological processes [7]. Next, the chapter introduces an exploratory variability analysis to rank brain regions by the stability of these spectral features. For each subject, we compute the coefficient of variation (CV) of each feature within each brain region over time and employ nonparametric statistical tests (Friedman tests with Kendall's W effect sizes) to assess whether certain regions (e.g., hippocampus, amygdala) exhibit significantly more variability than others. This analysis yields a patient-specific stability ranking of regions, highlighting, for instance, whether key limbic areas fluctuate more in their oscillatory dynamics compared to a reference region. We pay special attention to white matter as a reference: traditionally considered electrically quiescent, recent intracranial studies have shown that white matter contacts can display measurable oscillatory activity [8]. Chapter Three leverages white matter as a conservative baseline, recognizing that white matter contacts can produce measurable signals [8] and tests whether target regions show greater oscillatory variability than this baseline after controlling for factors such as the time of day. Finally, a Bayesian hierarchical model is constructed to formally test our primary hypotheses regarding differences in regional stability. This multilevel model accounts for the nested structure of the data (time windows within segments, segments within channels, channels within subjects) by including random effects that partially account for inter-subject and inter-electrode variability. Fixed effects (such as brain region and a day/night indicator for circadian timing) are then used to estimate differences in spectral feature stability between regions, with posterior credible intervals and probabilities providing direct evidence of effect significance. The Bayesian framework naturally propagates uncertainty at each level and allows us to incorporate the prior expectation that only large differences are scientifically noteworthy. We rigorously validate the model using posterior predictive checks and leave-one-out cross-validation to ensure that the findings are not artifacts of model misfit. Chapter Three's results demonstrate the effectiveness of this approach, as we find clear stratification of brain regions by signal stability. These differences persist even after accounting for diurnal variations, suggesting intrinsic physiological disparities in how these regions regulate neural oscillations. More broadly, Chapter Three illustrates a generalizable strategy for analyzing high-dimensional biomedical signals—one that emphasizes interpretable feature extraction, within-subject variability metrics, and hierarchical modeling to draw out biologically meaningful patterns from complex data.

Chapter Four focuses on a real-time monitoring application at the intersection of neurophysiology and clinical care: the intraoperative tracking of brain states in anesthetized infants and children. Titled "From Spectral Features to Perioperative Dynamics: Change Points Detection and Population-Averaged Modeling of Patient State Index," this chapter addresses the challenge of monitoring anesthetic depth in young patients using processed EEG signals. In pediatric anesthesia, traditional behavioral signs of depth are unreliable, so anesthesiologists increasingly rely on indices like the Patient State Index (PSI) – a derived 0–100 EEG-based score – to titrate anesthetic dosing [9, 10]. However, the pediatric brain is highly non-stationary: EEG readings can fluctuate abruptly due to arti-

facts (e.g., muscle movements, cardiogenic noise) or physiological events (e.g., cooling, bypass, burst-suppression) [11]. Simple averages of the PSI over time or across patients can, therefore, obscure critical shifts in brain state [11]. Chapter Four develops a novel statistical pipeline to capture and interpret these rapid neurophysiological transitions. We treat the second-by-second PSI time series as a stochastic process and explicitly model change points—moments when the underlying distribution of PSI values changes—as key indicators of brain state instability. At a macro scale, we apply the Pruned Exact Linear Time algorithm (PELT) to each patient's concatenated PSI trajectory, detecting abrupt changes aligned with surgical phases [12]. At a finer scale, we deploy a Bayesian structural time series (state-space) model for each patient, which infers latent shifts in the PSI level with full posterior uncertainty [13]. This two-tiered approach (external algorithmic segmentation and internal Bayesian filtering) yields a sequence of candidate change points along each patient's anesthetic course. We then introduce a phase-normalized metric of instability, the "Variability Ratio Index (VARI)," defined as the probability of a state transition within a given time window, adjusted for the length of each surgical phase. VARI provides a comparably scaled measure of how labile the PSI is during, say, induction versus maintenance or emergence. Finally, Chapter Four links these signal-derived events to clinical covariates through a population-level regression model. Specifically, we fit a logistic GEE that treats each one-second interval (nested within patient and surgical phase) as an observation of "instability" or "stability," with covariates including the current phase of surgery, patient demographics (age, sex, body mass index, and ethnicity), and other factors. The GEE uses a logit link and incorporates a custom working correlation (estimated residual autocorrelation from the data) to account for the serial dependence in these binary outcomes [14]. This yields odds ratio estimates for the effect of each covariates on the instantaneous probability of a PSI change point, along with robust standard errors for inference. Model diagnostics in Chapter Four confirm that this approach successfully captures the within-patient correlation structure and that the fitted probabilities are well calibrated to the observed data. The substantive findings offer new insights into pediatric anesthetic management: for example, the analysis reveals that surgical events and phase transitions are associated with sharp increases in PSI instability, whereas static patient characteristics, such as gender or broad ethnicity, have a negligible influence on these rapid EEG fluctuations. We also observe a significant reduction in instability with increasing patient age and higher body mass. One possible explanation is that more mature (older/heavier) toddlers have inherently more stable neurophysiology, though this remains speculative without direct physiological evidence. By translating raw EEG signals into a probability of state change, Chapter Four provides clinicians with an interpretable metric of how responsive or volatile a child's brain activity is under anesthesia. This approach thus enhances the monitoring of anesthetic depth beyond simple threshold-based indices, potentially helping to identify moments of inadequate or excessive anesthesia in real time. In sum, Chapter Four contributes a probability-first analytic framework for high-frequency physiologic data, demonstrating how blending algorithmic detection with statistical modeling can uncover clinically relevant patterns in complex perioperative signals.

The three research chapters are united by their goal of integrating diverse data streams to improve health monitoring and outcomes. Each chapter addresses a distinct scenario – from population-level surveillance of infectious disease outcomes (leveraging EHR and PRO data), to controlled research analysis of intracranial neural signals, to real-time moni-

toring of pediatric patients during surgery – yet all share common methodological threads. First, they all prioritize interpretability and scientific validity in the face of complex data: whether it is translating model outputs into easily understood effect sizes and probabilities (Chapters Two and Four) or extracting physiologically meaningful features from raw signals (Chapter Three), the emphasis is on making high-dimensional data comprehensible to clinicians and stakeholders. Second, a blend of Frequentist and Bayesian methods is used to ensure robust inference. For instance, Chapters Two and Four use classical regression frameworks (GLM, GEE) for estimation clarity while incorporating Bayesian-inspired elements (hierarchical structuring, prior-informed metrics) to handle multilevel or sparse information. In contrast, Chapter Three leans on Bayesian multilevel modeling but is informed by descriptive Frequentist tests and criteria for model checking. This blended approach reflects a guiding philosophy of the thesis: complex health data benefits from multiple analytical lenses, and embracing both paradigms allows one to cross-validate findings and capture different dimensions of uncertainty [3]. Third, each analysis explicitly models heterogeneity rather than treating it as a nuisance. The analyses explicitly model it—be it temporal fluctuations in PRO-related health behaviors, differences in neural signal stability across brain regions, or second-to-second volatility in anesthesia depth. By doing so, we can characterize when and why outcomes deviate from expectations (e.g., a sudden spike in antibiotic use, a transient neural oscillation surge, or an instability in the EEG index) and tie these deviations to underlying factors. Finally, rigorous validation and calibration procedures are hallmarks of the work presented in every chapter. Missing data are carefully imputed or sensitivity-tested (Chapter Two); model assumptions and fits are scrutinized via simulation-based diagnostics (Chapter Three), and predictive performance is evaluated against held-out or pseudo-prospective data (Chapters Two and Four). These measures ensure that the conclusions drawn are reliable and grounded in evidence rather than artifacts of any particular modeling choice.

Overall, this thesis demonstrates a coherent framework for monitoring patient-reported and physiological outcomes through integrated data analytics. By developing statistical methodologies that accommodate high-dimensional signals, align with clinical knowledge, and preserve interpretability, we contribute to the emerging paradigm of data-driven precision monitoring in medicine. The approaches detailed in Chapters Two, Three, and Four illustrate how diverse data (ranging from survey responses to EEG time-series) can be synthesized to glean actionable insights – from guiding public health interventions to informing bedside clinical decisions. Taken together, the work advances both the theory and application of Biostatistics in an era where rich data streams are increasingly available to inform patient care. The remainder of this dissertation delves into each study in detail, providing the specific context, methods, results, and implications that underpin this integrated vision of patient outcomes and neurophysiological monitoring.

Chapter 2

Classical and Bayesian Statistics in Infectious Disease Analysis and Surveillance

2.1 Introduction

Infectious disease surveillance depends on the timely collection, synthesis, and interpretation of health signals to guide prevention and control. Routine data streams, like case counts, hospital encounters, prescribing records, etc., exhibit trends, seasonality, and transient perturbations that call for formal time–series analysis [15]. In practice, translating these structures into actionable inferences requires models that can accommodate autocorrelation and nonstationarity while remaining interpretable to public health stakeholders [16]. Classical ARIMA-type specifications offer a transparent baseline for signal extraction, forecasting, and uncertainty quantification [17]. Emerging applications extend these ideas to real-time forecasting and early warning, where short-horizon predictions support situational awareness and resource allocation [18]. When policy shifts or communication campaigns occur, interrupted time-series (ITS) designs provide a principled framework for quantifying level and slope changes against an evolving background process [19]. Bayesian counterparts to these models further enable coherent propagation of uncertainty and incorporation of prior information, which can be especially valuable in sparse or rapidly evolving surveillance contexts [20].

Frequentist and Bayesian paradigms offer complementary perspectives for inference in infectious disease research. Frequentist procedures treat parameters as fixed but unknown, characterizing uncertainty through sampling distributions, confidence intervals, and likelihood-based tests, which yield estimates with well-defined long-run properties [21]. Bayesian analysis treats parameters as random variables, combines prior knowledge with observed data via Bayes' theorem, and returns posterior distributions that support direct probabilistic statements about effects and predictions [22]. Rather than being viewed as competitors, the two paradigms function as a coherent toolkit: Frequentist estimators and tests provide robust baselines and diagnostics, while Bayesian models facilitate the borrowing of information across strata, propagate uncertainty through hierarchical structures, and generate full predictive distributions for decision-making [21]. This dual perspective is particularly useful when integrating heterogeneous sources like clinical

records, patient-reported outcomes, and digital traces, where prior structures and explicit probability statements aid interpretation [22].

The analytical strategy in infectious disease analysis and surveillance spans models for independent outcomes, as well as extensions for dependence, prediction, and missingness. For mean structure and association, GLMs can be estimated using both Frequentist and Bayesian approaches. In high-dimensional settings, penalized regression methods (e.g., Ridge regression and LASSO) are used for shrinkage and variable selection. [23]. Model specification and selection combine likelihood ratio testing and information criteria with Bayesian approaches that prioritize out-of-sample predictive adequacy and stability; this balance supports parsimony without sacrificing interpretability [22]. For probabilistic prediction, Naive Bayes, Bayesian networks with data-driven structure and parameter learning, and Bayesian Additive Regression Trees capture nonlinearities and interactions while retaining well-calibrated uncertainty [21]. Correlation from clustering or longitudinal follow-up is addressed with GLMM, GEE, and Bayesian hierarchical models, enabling partial pooling and robust inference under complex sampling [22]. Diagnostic work evaluates residual structure, dispersion, and fit; for binary and count endpoints, while overdispersion is addressed through quasi-likelihood, negative binomial, and zero-inflated formulations. A fully Bayesian approach can incorporate prior information and provide full posterior uncertainty intervals, which are especially useful in small samples [17]. In policy evaluation and surveillance forecasting, the classical and Bayesian ITS models, including ARIMA and structural variants, are applied to estimate immediate and gradual intervention effects [19]. Finally, missing data are addressed within the missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) frameworks using multiple imputation by chained equations and fully Bayesian joint modeling, with congeniality between imputation and analysis models and full propagation of imputation uncertainty [24]. Sensitivity analyses to departures from MAR (e.g., delta-adjustment and pattern-mixture priors) accompany the primary results, yielding valid estimation and prediction under plausible mechanisms and aligning with established theory and applied guidance as well as recent Bayesian advances [25–28].

To demonstrate these methods on a realistic surveillance platform, we analyzed de-identified records from the All of Us Research Program's Registered Tier Curated Data Repository, which integrates participant surveys, electronic health records, physical measurements, biospecimens, and selected digital inputs in a secure cloud environment [29, 30]. We constructed two complementary datasets to mirror the modeling agenda above: (i) an individual-level cohort linking infectious-disease encounters to downstream antibiotic orders/fills, laboratory results, vitals, demographics, comorbidity burden, vaccination history, and prior utilization that support cross-sectional and longitudinal GLM/GLMM/GEE analyses, Bayesian hierarchical models, and predictive learners such as penalized regression and BART; and (ii) a surveillance series aggregating, by calendar month, antibiotic prescribing following viral diagnoses, expressed as rates per 1,000 relevant visits and stratified by site/region for interrupted time-series analyses with ARIMA and structural variants to evaluate a national stewardship campaign [19, 31]. The harmonized representation enables consistent feature engineering across health systems and justifies partial pooling by participant and care site in hierarchical specifications [22]. Missing covariates and outcomes are addressed under MCAR, MAR, and MNAR assumptions via multiple imputation by chained equations and fully Bayesian joint models, along

with sensitivity analyses for departures from MAR [24–28]. This multi-modal design aligns the empirical applications with core public-health questions on burden, behavior, and intervention response.

All analyses are conducted in R/RStudio to ensure reproducible workflows, literate programming, and tight integration between data processing and modeling [32]. Bayesian estimation leverages BUGS-family software i.e., WinBUGS, OpenBUGS, and JAGS for Gibbs sampling and model specification using declarative syntax [33]. The Stan ecosystem provides gradient-based sampling via HMC/NUTS, along with optimization and variational inference, accessed through `rstan`, `rstanarm`, and `brms` to streamline model fitting and diagnostics [34]. Interfaces maintained for OpenBUGS and `rjags` facilitate rapid prototyping and sensitivity analyses across engines [35]. High-level modeling with `brms` accelerates specification of multilevel and distributionally rich models while preserving access to Stan's inference back end [36]. This tool-chain supports time-series modeling, imputation, and infectious disease workflows within a unified environment, enabling transparent estimation, comprehensive diagnostics, and replicable reporting [37, 38].

2.2 Bayesian and Frequentist Approaches in Infectious Disease Data Analysis

Frequentist inference treats parameters as fixed but unknown and characterizes uncertainty through sampling distributions, confidence intervals, and likelihood-based tests; this perspective has long supported standardized reporting in epidemiology and routine surveillance where long-run operating characteristics are central [21]. Bayesian inference treats parameters as random variables and updates prior beliefs with the likelihood to obtain posterior distributions that permit probability statements about effects and predictions; this perspective aligns naturally with hierarchical structures, small strata, and sequential learning in public health [22]. In practice, the two perspectives are not adversarial but complementary: Frequentist estimators provide transparent baselines and calibration, whereas Bayesian models propagate uncertainty across levels and deliver predictive quantities that map directly to decision-making about interventions, burden, and communication effects [21].

The descriptive strengths of both perspectives converge at the point of scientific reporting: an effect size with principled uncertainty, framed against clinically meaningful thresholds and accompanied by diagnostics of model adequacy. In that spirit, the evidence summaries introduced in further subsections do not replace Frequentist quantities; rather, they translate posterior information into statements that are readily interpretable alongside familiar estimates and intervals, a particularly useful feature when integrating clinical records with patient-reported outcomes (PROs) and neurophysiological signals where heterogeneity and hierarchy are intrinsic [22].

2.2.1 Bayesian Evidence Summaries

The *probability of direction* (PD) is the posterior probability that an effect shares the sign of its posterior mean, thereby providing an immediate readout of directional evidence while remaining agnostic to the unit scale [39]. In early evaluation of stewardship or

communication campaigns, PD is often reported with the posterior median and interval so that a short-horizon decision about increase versus decrease can be articulated without overstating magnitude, PD has been particularly informative when the primary question is whether a change in symptom burden or activity is upward or downward following an intervention [40]. The *region of practical equivalence* (ROPE) complements the directional summary by formalizing negligible effects via a domain-specific interval around zero; the posterior mass within this interval operationalizes practical irrelevance and anchors inference to minimally important differences on the log-odds, risk, rate, or raw-score scales relevant to clinical and behavioral outcomes [41]. When explicit statements about a sharp null are warranted, Bayes factors via the Savage–Dickey density ratio quantify evidence by comparing prior and posterior density at the null value; this device is particularly appropriate in interrupted time-series settings where the absence of a level change is a meaningful claim, with interpretation presented alongside prior sensitivity analyses to reflect dependence on local prior behavior [42].

2.2.2 Prior Specification

Weakly informative priors are commonly adopted to stabilize estimation by gently regularizing extreme coefficients while allowing the likelihood to dominate as information accumulates; such priors are well matched to logistic and Poisson models where separation or sparse cells may arise, and they provide interpretable anchoring for effects that are expected to be small on substantive grounds [22]. Informative priors enter naturally when credible external evidence exists i.e., previously validated effects, baseline rates, or historical prevalence, and their influence is most visible in small-sample or subgroup contexts, so provenance and sensitivity to plausible alternatives are typically documented as part of transparent reporting [21]. When historical information is relevant and commensurate, formal borrowing has been adopted through power priors that discount historical likelihoods by a weight parameter, commensurate priors that adapt borrowing to empirical similarity, and meta-analytic-predictive (MAP) priors that summarize multiple studies under a random-effects structure; robust MAP variants introduce a weakly informative component to mitigate prior–data conflict in dynamic surveillance or multi-source integration [43]. In PRO applications, MAP priors have been particularly suitable for initializing baseline distributions of symptom scores or activity metrics from earlier cohorts before new follow-up accrues [22].

Structured expert elicitation has been widely used when empirical evidence is sparse or lagged, with quintile or probability judgments mapped to parametric families so that uncertainty is explicitly represented [44]. In infectious-disease analytics, this strategy has supported early assessments of severity and adherence behaviors; in PRO contexts, it has underpinned prior beliefs about minimally important differences or plausible effect sizes for behavioral interventions, with fitted priors compared against literature-informed alternatives to demonstrate robustness [45].

2.2.3 Posterior Computation and Diagnostics

Posterior inference is obtained with Hamiltonian Monte Carlo using the No-U-Turn Sampler, as implemented in the R ecosystem through `rstan` and the high-level interface `brms`;

this combination provides efficient gradient-based sampling for multilevel models while maintaining accessible specification and reproducibility [34]. Gibbs-sampling engines in the BUGS/JAGS family remain established for conditionally conjugate structures and for rapid prototyping, with mature interfaces that integrate into epidemiologic pipelines [33]. Convergence assessment is typically presented through multiple chains with visual checks (trace plots and marginal densities) together with scalar diagnostics such as effective sample size and the rank-normalized $\hat{R}$; model adequacy is often summarized with posterior predictive checks and information-focused criteria (PSIS-LOO or WAIC), reported alongside uncertainty summaries to align with transparent decision-making [46].

The paradigm framing, evidence summaries, and prior strategies provide a coherent pathway from heterogeneous data to decision-relevant inference. In surveillance applications where immediate direction is central, PD is reported to communicate the dominant sign of change, whereas ROPE anchors conclusions to clinically or behaviorally meaningful thresholds [40]. When the absence of an effect is the scientifically relevant assertion, Bayes factors based on Savage–Dickey have been preferred and are accompanied by sensitivity displays to clarify prior dependencies [42]. Across regression and time-series models, weakly informative priors stabilize estimation and preserve interpretability, while MAP or commensurate constructions transport credible historical information into current analyses of PROs in a manner that respects heterogeneity [43]. Computation and diagnostics close the loop by demonstrating that posterior summaries are stable and that model fit is adequate for the intended questions, thereby sustaining a standard presentation that integrates clinical records with digital and social media data [34].

2.3 Generalized Linear Models in Infectious Disease Analysis and Surveillance: Methods for Independent Data

We applied GLM on real-world data from the All of Us Research Program; here we detail the cohort construction and measure preparation that underlie those models. The study population comprised individuals with at least one qualifying hospital procedure, with the first qualifying encounter set as the index. The primary outcome was a binary indicator of infection-related disease occurring within 180 days post-index, identified from ICD-10-CM diagnoses under "Certain infectious and parasitic diseases."

Table 2.1 presents the baseline characteristics of the study cohort. The cohort comprises a total of 881 patients, of which 105 (approximately 11.9%) developed a post-procedural infection. Demographic variables were harmonized to three-level categories for race (White; Black or African American; Other) and ethnicity (Hispanic or Latino; Not Hispanic or Latino; Other), alongside sex at birth (Female; Male; Other). Clinical covariates captured pre-existing comorbidities spanning circulatory, respiratory, endocrine/metabolic, neoplasms (cancer), and a residual "other disorders" group; anthropometric included body mass index (BMI) as a continuous measure summarized by median and quartiles. Procedure text strings were mapped into clinically interpretable groups—catheter/drainage/dialysis, endoscopy, intubation, biopsy, transplant, replacement, and surgery—then encoded as binary indicators for analysis. To balance computational efficiency and representativeness, a simple random 10% sample of eligible participants was drawn, and two analytic representations were created: a wide format (one

Table 2.1: Baseline Characteristics by Infection Status

Characteristic	**No (N = 776)**	**Yes (N = 105)**
Age Median (Q1, Q3)	52 (39, 62)	55 (45, 65)
Sex at Birth		
Female	471 (61%)	52 (50%)
Male	292 (38%)	52 (50%)
Other	13 (1.7%)	1 (1.0%)
Race		
Black or African American	162 (21%)	24 (23%)
Other	239 (31%)	37 (35%)
White	375 (48%)	44 (42%)
Ethnicity		
Hispanic or Latino	155 (20%)	26 (25%)
Not Hispanic or Latino	591 (76%)	77 (73%)
Other	30 (3.9%)	2 (1.9%)
BMI Median (Q1, Q3)	30 (26, 36)	29 (26, 36)
Previous Infection n (%)	155 (20%)	33 (31%)
Circulatory Disorder n (%)	30 (3.9%)	14 (13%)
Respiratory Disorder n (%)	19 (2.4%)	11 (10%)
Endocrine Disorder n (%)	116 (15%)	40 (38%)
Cancer n (%)	90 (12%)	20 (19%)
Other Disorders n (%)	12 (1.5%)	3 (2.9%)
Comorbidity (2+) n (%)	61 (7.9%)	20 (19%)
Procedural Groups		
Catheter/Drainage/Dialysis n (%)	30 (3.9%)	8 (7.6%)
Endoscopy n (%)	373 (48%)	31 (30%)
Intubation n (%)	34 (4.4%)	22 (21%)
Biopsy n (%)	156 (20%)	16 (15%)
Transplant n (%)	19 (2.4%)	8 (7.6%)
Replacement n (%)	29 (3.7%)	1 (1.0%)
Surgery n (%)	99 (13%)	13 (12%)

row per participant) for independent-data models and a long format (repeated procedures across yearly intervals) to support longitudinal specifications. Baseline characteristics indicated older age among those with subsequent infections (median 55 [45–65] years versus 52 [39–62] years), a higher prevalence of endocrine disorders (38% vs. 15%), and distinct procedural profiles, with greater frequencies of intubation (21% vs. 4.4%) and transplant (7.6% vs. 2.4%) among infected patients, while endoscopy was more common in those without infections (48% vs. 30%).

2.3.1 Bayesian GLM

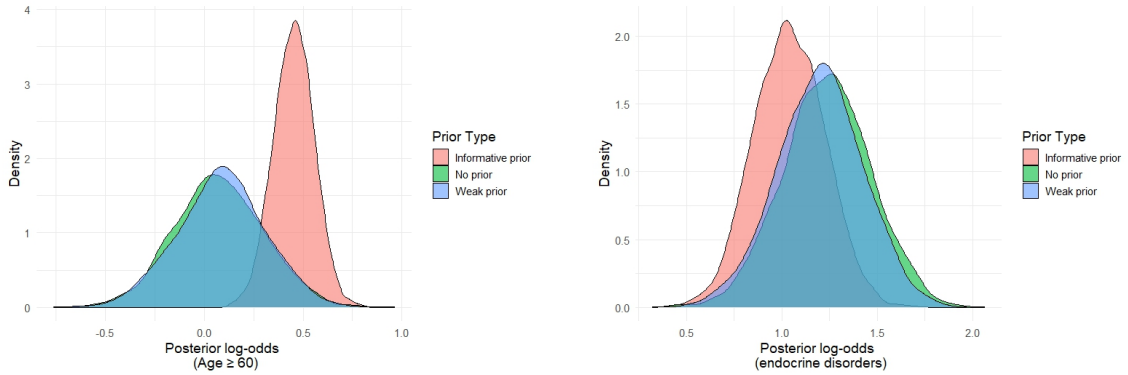
We applied GLM on the All of Us cohort in a Bayesian framework with the specific intention of illustrating how posterior inference changes under alternative prior specifications, holding the likelihood and outcome definition fixed. Three logistic models (Bayesian Generalized Linear models denoted Bglm.1 - Bglm.3) were estimated for the binary endpoint (infection within 180 days): *Bglm.1* with non-informative priors, *Bglm.2* with weakly informative priors, and *Bglm.3* with informative priors calibrated from external evidence for age $\geq$ 60 and endocrine disease (meta-analytic odds ratios transformed to the log-odds scale, with standard deviations obtained from confidence limits). Model fitting relied on Hamiltonian Monte Carlo with the No-U-Turn Sampler to ensure efficient exploration of the posterior, and reporting focused on odds ratios (ORs) with 95% credible intervals for interpretability consistent with epidemiologic practice [47, 48]. The prior construction followed a principled pathway in which external evidence aligned with plausible clinical directionality for age and endocrine conditions, while leaving other coefficients minimally constrained; computation and convergence were supported by gradient-based sampling and standard diagnostics for modern Bayesian GLMs [49, 50]. The resulting summaries are presented in Table 2.2, with posterior distribution plots for the age and endocrine effects in Figures 2.1, to make the effect of prior information visually explicit [51, 52].

Table 2.2: Estimated Odds Ratios (OR) and 95% Credible Intervals from Bayesian GLMs

Variable	**Level**	**BGLM 1**	**BGLM 2**	**BGLM 3**
Intercept	–	0.10 (0.07, 0.13)	0.10 (0.07, 0.13)	0.09 (0.07, 0.11)
Age 60+	No	Ref	Ref	Ref
	Yes	1.08 (0.70, 1.63)	1.09 (0.70, 1.69)	1.57 (1.29, 1.93)
Endocrine Disorder	No	Ref	Ref	Ref
	Yes	3.46 (2.24, 5.40)	3.35 (2.12, 5.24)	2.82 (1.97, 4.08)

The posterior patterns in Table 2.2 and Figures 2.1 show a coherent progression as prior information becomes more specific. Under non-informative and weakly informative priors (BGLM 1–2), the age effect centers near the null with wide intervals [OR 1.08 (0.70–1.63) and 1.09 (0.70–1.69)], reflecting limited information in the data alone to resolve a moderate association for the 60+ indicator; the weak prior produces negligible movement, consistent with its stabilizing rather than assertive role [47]. With informative priors aligned to published evidence, the age 60+ effect shifts upward and the interval contracts [OR 1.57 (1.29–1.93)], an expected pattern when external knowledge is commensurate with the observed cohort and the posterior synthesizes both sources [51]. For endocrine disease,

the posterior already indicates a strong positive association under minimal prior input [OR 3.46 (2.24–5.40) in BGLM 1], remains similar with weak priors [3.35 (2.12–5.24)], and settles to a slightly smaller yet still pronounced effect under the informative prior [2.82 (1.97–4.08)], a trajectory that is consistent with prior centering nearer to moderate effects while preserving the substantive signal in the data. The intercept shows modest attenuation in BGLM 3 [0.09 (0.07–0.11) vs. 0.10 (0.07–0.13)], reflecting the integrated baseline risk after prior incorporation.



(a) Posterior Distribution of Age Effect.

(b) Posterior Distribution of Endocrine Disease Effect.

Figure 2.1: Posterior distributions of covariates effects.

These contrasts illustrate the intended message: informative priors sharpen age-related inference where data are relatively less decisive, while endocrine disease remains a robust predictor across specifications; the density plots in Figures 2.1 visualize these differences as shifts in location and reductions in spread, and the OR scale anchors interpretation in a clinically familiar metric for post-procedural infection risk [49, 52].

2.3.2 Frequentist GLM

Within the Frequentist framework, logistic GLM were estimated by maximum likelihood to quantify the association between infection within 180 days and two clinically motivated predictors (age $\geq$ 60 and endocrine disorder), with effects reported as odds ratios (OR) and 95% confidence intervals; this presentation aligns with standard epidemiological reporting and the likelihood-based fitting of GLM [47, 48].

Table 2.3: Odds Ratios with 95% Confidence Intervals and P-values from Frequentist Logistic Regression

Variable	**Odds Ratio**	**95% CI (Lower)**	**95% CI (Upper)**	**P-value**
Intercept	0.10	0.07	0.13	< 0.001
Age 60+	1.09	0.69	1.68	0.71
Endocrine Disorder	3.44	2.18	5.38	< 0.001

The main–effects specification (Table 2.3) indicates a precise and substantial elevation in risk for endocrine disorder [OR 3.44, 95% CI 2.18–5.38; $p < 0.001$], whereas the age indicator centers near the null with wide uncertainty [OR 1.09, 95% CI 0.69–1.68; $p = 0.71$], consistent with an imprecise estimate under the additive model [52]. Extending the model with an interaction term (Table 2.4) leaves the age main effect non-significant [OR 1.36, 95% CI 0.78–2.32; $p = 0.26$] and further accentuates the endocrine association [OR 4.47, 95% CI 2.46–7.96; $p < 0.001$]; the interaction coefficient itself suggests attenuation of the endocrine effect among older adults [OR 0.54, 95% CI 0.22–1.34; $p = 0.18$] but does not reach conventional significance.

Table 2.4: Odds Ratios with 95% Confidence Intervals and P-values from Frequentist Logistic Regression with Interaction Term

Variable	**Odds Ratio**	**95% CI (Lower)**	**95% CI (Upper)**	**P-value**
Intercept	0.09	0.07	0.12	< 0.001
Age 60+	1.36	0.78	2.32	0.2609
Endocrine Disorder	4.47	2.46	7.96	< 0.001
Age 60+ $\times$ Endocrine Disorder	0.54	0.22	1.34	0.1817

Predicted probabilities in Figure 2.2 provide a complementary display: within the < 60 group, endocrine disorder corresponds to markedly higher risk (approximately 28% vs. 8%), whereas among those ≥ 60 the contrast narrows (approximately 24% vs. 11%), yielding non-parallel profiles that visually echo the non-significant interaction [53].

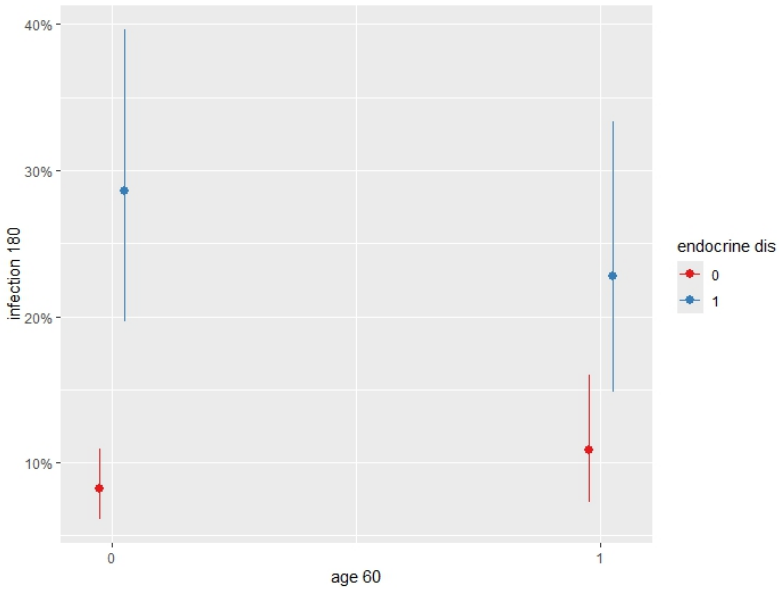


Figure 2.2: Interaction Plot between Age and Endocrine Disorders.

2.3.3 Penalized Regression

Penalized logistic regression was estimated to address collinearity among clinical and procedural covariates while maintaining an interpretable odds–ratio scale. LASSO (ℓ_1) and Ridge (ℓ_2) penalties were fitted with `glmnet`, and the regularization parameter λ was selected by 10-fold cross–validation to balance fit and parsimony [54, 55]. For Ridge, the cross–validation curve identified $\log(\lambda_{\text{min}}) = -2.34$ as the deviance–minimizing value and $\log(\lambda_{1\text{se}}) = 3.8$ as the more conservative choice within one standard error; as expected with an ℓ_2 penalty, coefficients remained non–zero across the path (Figure 2.3).

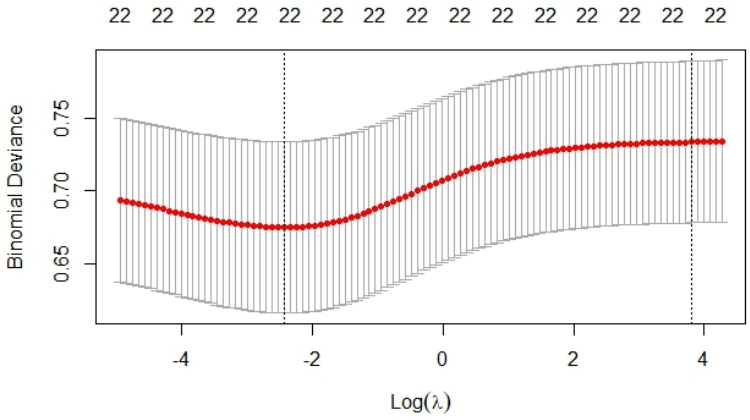


Figure 2.3: 10-fold cross-validation curve for Ridge regression model.

For LASSO, $\log(\lambda_{\text{min}}) = -4.5$ achieved the lowest cross–validated deviance, whereas $\log(\lambda_{1\text{se}}) = -2.6$ produced a sparse solution with only two non–zero effects, illustrating the selection property of the ℓ_1 penalty (Figure 2.4).

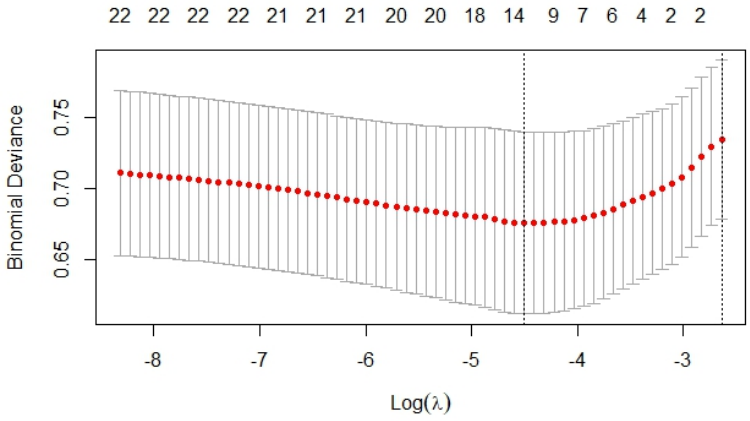


Figure 2.4: 10-fold cross-validation curve for LASSO regression model.

Because shrinkage induces bias, post–selection (debiased) inference was obtained with the `hdi` procedures `ridge.proj` and `lasso.proj`, yielding adjusted odds ratios

(OR_{adj}) and confidence intervals with improved inferential validity in high–dimensional or collinear settings [56–58]. On the All of Us cohort, the Ridge results (Table 2.5) indicated statistically credible associations (after debiasing) for circulatory (OR_{adj} 1.20, 95% CI 1.06–1.36), respiratory (1.29, 1.13–1.47), endocrine (1.17, 1.09–1.26), and cancer (1.09, 1.01–1.18) disorders, with intubation (1.25, 1.14–1.38) and transplant (1.19, 1.05–1.35) also elevated; multi–comorbidity showed an inverse adjusted association (0.82, 0.72–0.94), consistent with coding and covariates adjustment used in the model.

Table 2.5: Ridge regression: Odds Ratios (OR) and debiased OR (OR_{adj}) with 95% CI

Fixed Effect	**OR**	**OR_{adj}**	**Lower 95% CI**	**Upper 95% CI**
Age	1.01	1.00	1.00	1.00
Sex at birth: Male	1.20	1.02	0.98	1.07
Sex at birth: Other	0.73	0.93	0.79	1.11
Race: Other	1.07	0.99	0.91	1.08
Race: White	0.87	0.97	0.92	1.03
Ethnicity: Not Hispanic	0.92	0.98	0.90	1.06
Ethnicity: Other	0.67	0.92	0.81	1.04
Body Mass Index (BMI)	1.00	1.00	1.00	1.00
Previous infection	1.02	1.00	0.99	1.01
Circulatory disorder	1.67	1.20	1.06	1.36
Respiratory disorder	2.35	1.29	1.13	1.47
Endocrine disorder	1.80	1.17	1.09	1.26
Cancer	1.27	1.09	1.01	1.18
Other disorders	1.32	1.12	0.94	1.35
Comorbidity (2+)	0.93	0.82	0.72	0.94
Catheter/Drainage/Dialysis	1.48	1.07	0.96	1.19
Endoscopy	0.76	0.97	0.92	1.01
Intubation	2.65	1.25	1.14	1.38
Biopsy	0.90	0.99	0.93	1.05
Transplant	2.12	1.19	1.05	1.35
Replacement	0.65	0.95	0.84	1.07
Surgery	1.05	1.02	0.96	1.09

LASSO produced a concordant pattern (Table 2.6): among the 13 variables selected at λ_{min}, adjusted effects greater than one were observed for circulatory, respiratory, endocrine, and cancer disorders, as well as for intubation and transplant, while endoscopy and replacement trended below or near unity, aligning with their baseline distributions.

These penalized estimates preserve the substantive ranking of predictors, particularly the respiratory and endocrine signals and the procedural risks of intubation and transplant, while the cross–validation diagnostics make transparent the complexity–performance trade–off that underlies the chosen λ values and the resulting degree of shrinkage.

Table 2.6: Odds Ratios (OR) and debiased adjusted OR (OR_{adj}) from LASSO regression

Fixed Effect	**OR**	**OR_{adj}**	**Lower 95% CI**	**Upper 95% CI**
Age	1.00	1.00	0.999	1.002
Sex at birth: Male	1.16	1.03	0.979	1.073
Race: White	0.96	0.97	0.917	1.029
Previous infection	1.01	1.00	0.995	1.012
Circulatory disorder	1.51	1.20	1.064	1.362
Respiratory disorder	2.36	1.27	1.132	1.467
Endocrine disorder	2.20	1.17	1.092	1.256
Cancer	1.08	1.07	1.011	1.181
Catheter/Drainage/Dialysis	1.31	1.08	0.957	1.193
Endoscopy	0.82	0.96	0.921	1.014
Intubation	3.48	1.27	1.140	1.381
Transplant	2.37	1.19	1.046	1.353
Replacement	0.97	0.92	0.835	1.071

2.3.4 Bayesian Penalized Regression with Shrinkage Priors

Extending the penalized Frequentist analysis, Bayesian logistic regression was estimated under Gaussian (Ridge) and Laplace (LASSO) shrinkage priors to regularize coefficients while propagating full posterior uncertainty. In the Bayesian formulation, penalties enter as priors centered at zero—normally distributed coefficients induce the L_2 (Ridge) effect through a prior variance that governs global shrinkage, whereas Laplace (double–exponential) coefficients correspond to an L_1 (LASSO) effect with a sharp peak at zero and heavier tails that preferentially shrink small signals toward zero while allowing larger effects to escape shrinkage. Hyperparameter can be learned from the data, and the output includes posterior OR with credible intervals and an importance ranking, yielding an interpretable summary of signal strength under each prior. These models were fit with `bayesreg` (logistic link), drawing sufficiently long MCMC chains and summarizing results as median OR, 95% credible intervals, approximate t-statistics, and ranks.

Across both shrinkage priors, the clinical picture was stable. Intubation, endocrine disorders, and respiratory disorders consistently occupied the top ranks, with median OR well above 1 and credible intervals excluding the null. Table 2.7 reports the results from Bayesian logistic regression models with Ridge shrinkage priors. The median OR were 3.01 for intubation, 1.96 for endocrine disorders, and 2.51 for respiratory disorders; all three featured high importance ranks (1–3) and posterior intervals not spanning 1, indicating robust positive associations with infection within 180 days. Transplantation showed a moderate positive association (median OR $\approx$ 2.35) with wider uncertainty that straddled conventional thresholds, whereas endoscopy trended below 1 (median OR $\approx$ 0.74), consistent with a protective pattern though with uncertainty. Sociodemographic factors (sex at birth, race, ethnicity), BMI, prior infection, age, biopsy, and surgery clustered at lower ranks with posterior intervals generally overlapping 1, indicating weak or uncertain associations after shrinkage. The intercept corresponded to a baseline odds near 0.10, coherent with the observed low event frequency in the cohort.

Table 2.7: Bayesian Logistic Ridge Regression Results

Variable	**Median OR**	**95% Credible Interval**	**t-Statistic**	**Rank**
Intubation	3.01	[1.61, 5.68]	3.49	1
Endocrine disorder	1.96	[1.29, 3.11]	3.04	2
Respiratory disorder	2.51	[1.13, 5.47]	2.26	3
Transplant	2.35	[0.95, 5.52]	1.94	4
Endoscopy	0.74	[0.50, 1.07]	-1.60	5
Circulatory disorder	1.71	[0.86, 3.34]	1.51	6
Sex at birth: Male	1.26	[0.89, 1.82]	1.27	7
Ethnicity: Other	0.58	[0.18, 1.52]	-1.07	8
Cancer disorder	1.32	[0.79, 2.26]	1.07	8
Catheter/Drainage/Dialysis	1.54	[0.68, 3.32]	1.08	8
Replacement	0.58	[0.16, 1.62]	-0.97	8
Race: White	0.85	[0.57, 1.26]	-0.80	12
Sex at birth: Other	0.64	[0.12, 2.39]	-0.66	13
Race: Other	1.06	[0.67, 1.64]	0.26	13
Ethnicity: Not Hispanic	0.91	[0.55, 1.46]	-0.40	13
BMI	1.00	[0.98, 1.02]	-0.25	13
Previous infection	1.02	[0.96, 1.08]	0.70	13
Other disorder	1.24	[0.33, 3.96]	0.31	13
Age	1.08	[0.74, 1.60]	0.39	13
Biopsy	0.89	[0.56, 1.40]	-0.51	13
Surgery	1.07	[0.63, 1.78]	0.25	13
Intercept	0.10	[0.04, 0.27]	–	–

Table 2.8: Bayesian Logistic LASSO Regression Results

Variable	**Median OR**	**95% Credible Interval**	**t-Statistic**	**Rank**
Endocrine disorder	2.14	[1.27, 3.53]	2.91	1
Intubation	3.30	[1.59, 6.53]	3.33	1
Respiratory disorder	2.34	[1.03, 5.52]	1.94	3
Transplant	2.23	[0.92, 5.87]	1.63	4
Endoscopy	0.78	[0.49, 1.09]	-1.22	5
Sex at birth: Male	1.19	[0.89, 1.80]	1.05	6
Circulatory disorder	1.51	[0.84, 3.21]	1.26	6
Ethnicity: Other	0.70	[0.19, 1.48]	-0.84	8
Catheter/Drainage/Dialysis	1.38	[0.72, 3.22]	0.92	8
Sex at birth: Other	0.75	[0.13, 2.19]	-0.54	10
Race: White	0.91	[0.58, 1.21]	-0.66	10
Ethnicity: Not Hispanic	0.95	[0.60, 1.37]	-0.34	10
Cancer disorder	1.20	[0.81, 2.10]	0.86	10
Replacement	0.72	[0.16, 1.54]	-0.77	10
Race: Other	1.04	[0.73, 1.56]	0.29	15
BMI	1.00	[0.98, 1.01]	-0.25	15
Previous infection	1.01	[0.97, 1.07]	0.64	15
Other disorder	1.14	[0.40, 3.52]	0.28	15
Age	1.05	[0.76, 1.54]	0.30	15
Biopsy	0.94	[0.58, 1.34]	-0.39	15
Surgery	1.04	[0.67, 1.71]	0.25	15
Intercept	0.10	[0.05, 0.22]	–	–

Results under the Bayesian LASSO mirrored the Ridge findings. Table 2.8 reports the results from Bayesian logistic regression models with LASSO shrinkage priors. Intubation (median OR 3.30) and endocrine disorders (median OR 2.14) again led the ranking, followed by respiratory disorders (median OR 2.34) and transplant (median OR 2.23); endoscopy remained below 1 (median OR 0.78) with modest evidence, and the same set of demographic and ancillary clinical covariates showed limited contribution, with credible intervals typically spanning 1. Age 60+ remained near the null (median OR 1.05; rank 15). As expected, the Bayesian LASSO yielded continuous (nonzero) posteriors rather than exact zeros, providing coherent uncertainty quantification for all coefficients. So in Bayesian penalized regression, the Ridge and LASSO shrinkage priors delivered highly concordant signals on the principal risk factors (intubation, endocrine and respiratory disorders) while tempering small, noisy effects—an outcome that aligns with the conceptual role of shrinkage in high-parameter logistic models.

2.4 Variable Selection in Generalized Linear Models

Building on shrinkage-based estimation, we next emphasized principled model selection and comparison to balance predictive performance with parsimony, using the same All of Us cohort and outcome definition described earlier. In this section, the workflow proceeds in a deliberately parallel fashion across paradigms so that conclusions can be read on a common scale. On the Frequentist side, GLM are estimated by maximum likelihood, with candidate specifications organized through clinically motivated block entry and evaluated by likelihood–ratio tests and information criteria; stability checks for collinearity and separation are performed before settling on a compact specification, and effects are reported as odds ratios with confidence intervals and p-values in a format familiar for epidemiological interpretation. The Bayesian strand mirrors these goals but expresses model adequacy through out-of-sample predictive performance. To quantify variable-level evidence beyond a single winning model, Bayesian model averaging via Bayesian Adaptive Sampling aggregates support over many plausible GLMs; cumulative model probabilities and model-complexity profiles show how posterior mass concentrates on a small subset, while posterior inclusion probabilities provide an interpretable ranking of predictors that can be read alongside the Frequentist estimates [59]. Finally, projection predictive selection takes the best predictive Bayesian reference and derives reduced submodels whose accuracy and AUC remain close to the full model, with a cross-validated inclusion heatmap highlighting the order and stability with which variables enter; this closes the loop from estimation to selection by presenting a parsimonious specification that preserves predictive content while improving interpretability [49, 50].

2.4.1 Frequentist Variable Selection: Stepwise Regression

Building on the same All of Us cohort and binary infection outcome defined earlier, we used stepwise logistic regression as a pragmatic screen to simplify the model, recognizing that this approach can bias estimates and may overfit. To mitigate these issues, we also compared results to penalized (e.g., LASSO) and Bayesian selection methods and incorporated subject-matter knowledge into model building. The starting point was a full multivariable specification, including demographic factors, comorbidities, and recent

invasive procedures; candidate moves (additions and deletions) were evaluated by Akaike's Information Criterion (AIC), with the algorithm iterating until no further improvement could be achieved [60, 61]. Although stepwise procedures should be applied with caution and ideally complemented by domain knowledge and out-of-sample checks, they remain a pragmatic screening device in routine epidemiological modeling when the aim is clarity and parsimony [62].

Table 2.9: Final Model Predictors Selected via Stepwise Regression

Variable	Description
sex_at_birthMale	Male vs. female sex at birth
raceWhite	White vs. other racial groups
ethnicityOther	Other ethnicity (vs. reference)
respiratory_dis	Presence of respiratory disease
endocrine_dis	Presence of endocrine disease
cancer_dis	Presence of cancer
cathet_drainage_dialysis	Catheterization, drainage, or dialysis procedure
intubation	Underwent intubation
transplant	History of organ transplant

The final stepwise model achieved the minimum AIC (AIC = 584.93) and retained nine predictors (Table 2.9), balancing statistical fit with clinical face validity. These terms encompass respiratory and endocrine disorders and cancer, together with procedure-related exposures (intubation, transplant, catheterization/drainage/dialysis), while basic demographics did not contribute materially to AIC once clinical covariates were present. Adjusted associations expressed on the odds-ratio scale are reported in Table 2.10. Intubation showed the largest effect (OR = 4.98, 95% CI: 2.65–9.37), followed by transplant (OR = 4.29, 95% CI: 1.74–10.57), with substantial elevations also for respiratory (OR = 3.29, 95% CI: 1.40–7.76) and endocrine disease (OR = 2.67, 95% CI: 1.66–4.30). Cancer displayed a positive but imprecise association whose interval crossed unity. Catheterization/drainage/dialysis showed a positive trend (OR $\approx$ 1.77) but did not reach conventional significance (two-sided p $\approx$ 0.063), suggesting weak evidence for an association. This avoids implying a sharp cutoff. In contrast, sex at birth, race, and ethnicity were not statistically significant once clinical factors were accounted for, a pattern consistent with the broader multivariable analyses in this cohort. These stepwise results align with clinical expectations that invasive airway management and immunomodulating conditions mark a higher near-term infection risk after procedures, while highlighting where uncertainty remains for future confirmatory work [61, 62].

2.4.2 Bayesian Variable Selection

Bayesian variable selection offers a principled alternative to frequentist approaches, moving beyond point estimates and single-model decisions. Instead of relying solely on information criteria such as AIC or BIC, the Bayesian paradigm incorporates model uncertainty and evaluates predictive performance more holistically. Three widely used strategies in this context are Leave-One-Out Cross-Validation (LOO-CV), Bayesian Model Averaging (BMA), and Projection Predictive feature selection.

Table 2.10: Odds Ratios (OR) and 95% Confidence Intervals from Final Logistic Regression Model (Stepwise Selection)

Variable	OR	95% CI Lower	95% CI Upper
Sex at birth: Male	1.50	0.97	2.27
Race: White	0.68	0.43	1.06
Ethnicity: Other	0.36	0.08	1.34
Respiratory disorder	3.29	1.40	7.76
Endocrine disorder	2.67	1.66	4.30
Cancer	1.64	0.90	2.64
Cathet/Drainage/Dialysis	2.26	0.98	5.28
Intubation	4.98	2.65	9.37
Transplant	4.29	1.74	10.57

Leave–One–Out Cross–Validation (LOO–CV)

Following the stepwise Frequentist screening, Bayesian comparison focused on predictive adequacy under three prespecified GLMs—BGLM 1 (age $\geq$ 60 only), BGLM 2 (additive age $\geq$ 60 + endocrine disorder), and BGLM 3 (age-by-endocrine interaction)—evaluated with Pareto–smoothed importance sampling leave–one–out cross–validation (PSIS–LOO). Reliability of the importance weights was assessed via Pareto-k diagnostics; the three panels in Figures 2.5–2.7 show k values comfortably below the conventional thresholds, supporting the accuracy of the PSIS approximation and allowing ELPD_{LOO}, pLOO, and LOOIC to be read without corrective refits [46, 63].

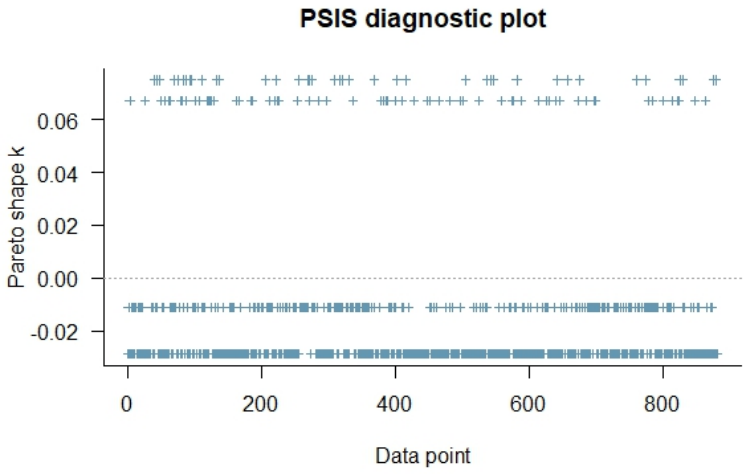


Figure 2.5: Pareto K diagnostics - Bayesian Model 1.

Quantitatively (Table 2.11), BGLM 2 attains the highest predictive score (ELPD_{LOO} = −310.65, SE = 19.18) and the lowest LOOIC (= 621.29, SE = 38.36), improving materially over the age-only baseline BM 1 (ELPD_{LOO} = −322.82; LOOIC = 645.65). BGLM 3 performs nearly identically to BGLM 2 (ELPD_{LOO} = −310.93; LOOIC = 621.86) but with larger effective complexity (pLOO = 4.32 vs. 3.17), yielding no discernible predictive gain once complexity is accounted for.

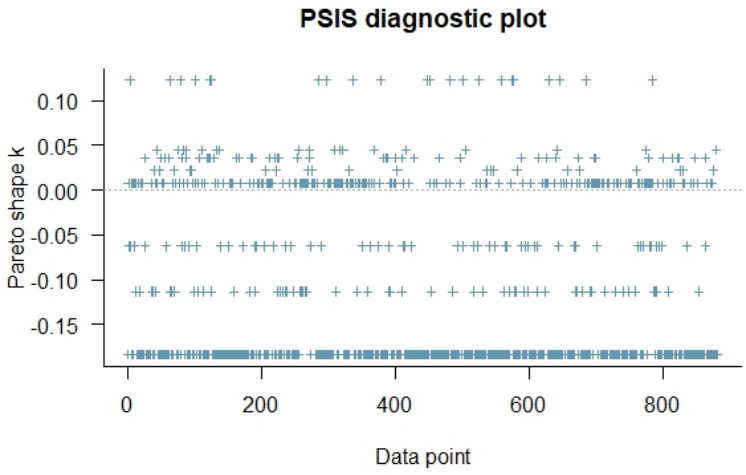


Figure 2.6: Pareto K diagnostics - Bayesian Model 2.

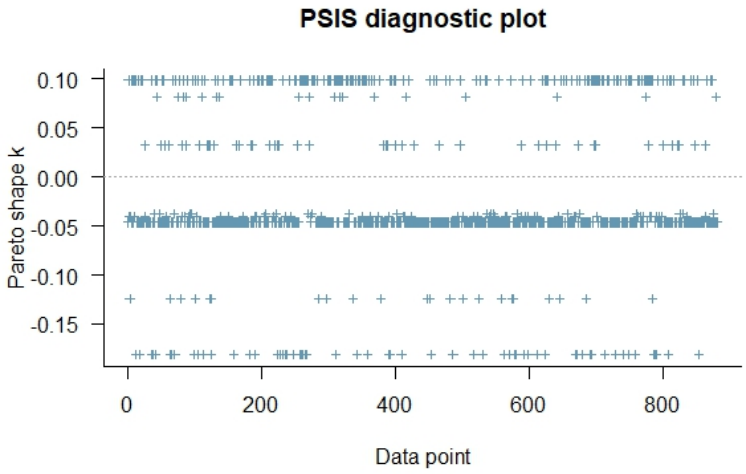


Figure 2.7: Pareto K diagnostics - Bayesian Model 3.

Table 2.11: Comparison of Bayesian Model (BM) Performance using LOO–CV Estimates

Metric	**BM 1: Age only**	**BM 2: Additive**	**BM 3: Interaction**
ELPD_{LOO}	−322.82 (SE = 19.36)	−310.65 (SE = 19.18)	−310.93 (SE = 19.25)
pLOO	2.03 (SE = 0.16)	3.17 (SE = 0.25)	4.32 (SE = 0.34)
LOOIC	645.65 (SE = 38.72)	621.29 (SE = 38.36)	621.86 (SE = 38.51)

Read alongside the stepwise results, the picture is coherent: an additive specification captures the dominant signal (endocrine disorder alongside age) with better out-of-sample fit than the baseline, and the interaction does not meaningfully improve predictive performance at the cohort scale considered here. These findings establish BGLM 2 as a well-calibrated reference for subsequent Bayesian averaging and projection, while the favorable Pareto-k profile documents that conclusions are not artifacts of unstable importance weights [49, 50, 63].

Model Averaging using Bayesian Adaptive Sampling

Following the LOO–CV comparison that favored the additive specification as a reference, model uncertainty was quantified through Bayesian model averaging (BMA) implemented via Bayesian Adaptive Sampling (BAS). In this framework, posterior model probabilities are estimated over a large space of GLM, and quantities of interest are summarized both at the model level and at the variable level. The cumulative model probability curve in Figure 2.8 rises steeply and plateaus early, indicating that a relatively small subset of candidate models captures most of the posterior mass; this pattern is consistent with a concentrated model space in which a few clinically coherent combinations dominate the evidence [47].

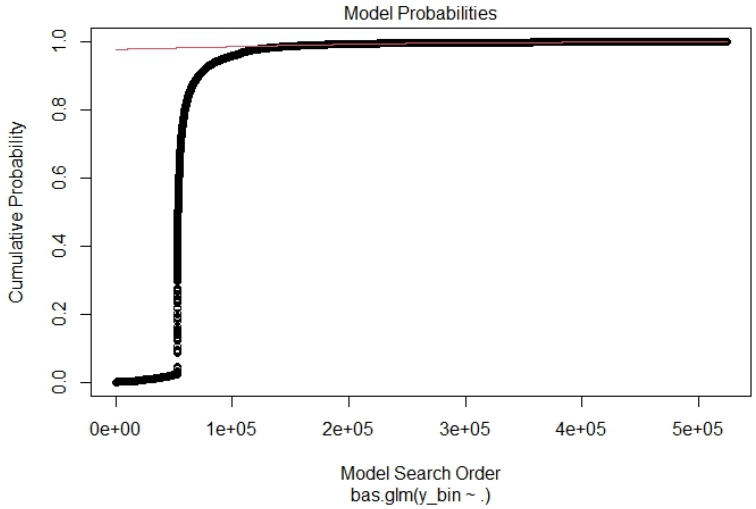


Figure 2.8: Cumulative Model Probabilities as a function of the model search order in BAS.

Complementing this view, the model–complexity profile in Figure 3.6 shows strongest support for moderate–sized specifications, with diminishing marginal gains as additional terms are introduced—an empirical expression of the accuracy–parsimony balance that also underpinned the Frequentist selection sequence [59].

Variable–level evidence summarized through posterior inclusion probabilities (PIPs) appears in Table 2.12. Intubation and endocrine disorder approach certainty (PIP $\approx$ 1), respiratory disorder and transplant carry substantial support (PIP $\approx$ 0.70), and other predictors show markedly lower inclusion evidence, aligning with the signals identified by stepwise screening and with the additive reference that maximized out–of–sample fit.

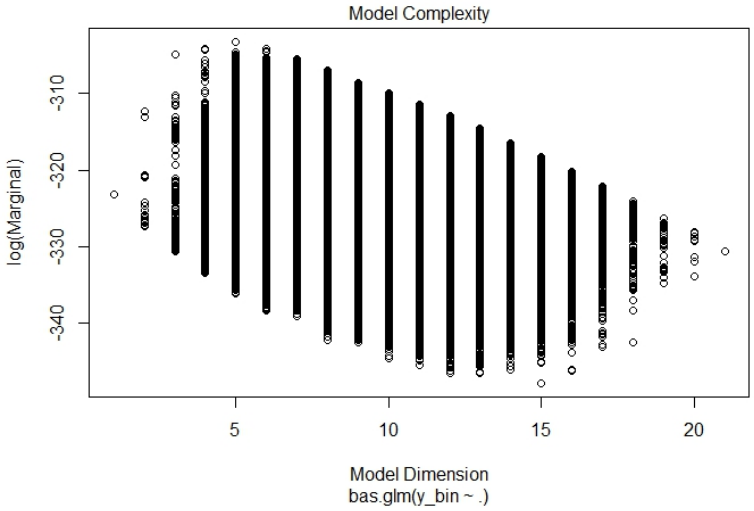


Figure 2.9: Model Complexity in BAS.

Table 2.12: Posterior Inclusion Probabilities (PIP) from BMA

Variable	**PIP**
Intercept	1.0000
Sex at birth: Male	0.2698
Circulatory disorder	0.2585
Respiratory disorder	0.7043
Endocrine disorder	0.9963
Catheter/Drainage/Dialysis	0.2244
Endoscopy	0.3018
Intubation	0.9965
Transplant	0.6957

Figures 2.8–2.9 and Table 2.12 indicate that the cohort's predictive structure is driven by a compact core of clinically plausible variables, while broader model uncertainty is well accounted for by averaging rather than by reliance on a single selected specification [47, 59].

Projection Predictive Feature Selection

With the additive Bayesian GLM (BGLM 2) established as the reference on the basis of LOO–CV, projection predictive selection was used to derive reduced sub-models that preserve the reference model's predictive distribution while improving parsimony. The sub-models were evaluated under 10-fold cross-validation, and performance was summarized as the difference in accuracy and AUC relative to the full reference. As shown in Figure 2.10, accuracy gaps diminish rapidly as the first few predictors are added and become negligible beyond a small set; Figure 2.11 shows the same pattern for AUC, with curves flattening once the core clinical signals are included. The cross-validated inclusion heatmap in Figure 2.12 clarifies both order and stability: intubation and endocrine disorder enter first and appear in essentially all folds; respiratory disorder and transplant follow closely and stabilize by modest sub-model sizes; sex at birth and endoscopy round out a compact specification whose discrimination remains close to the reference.

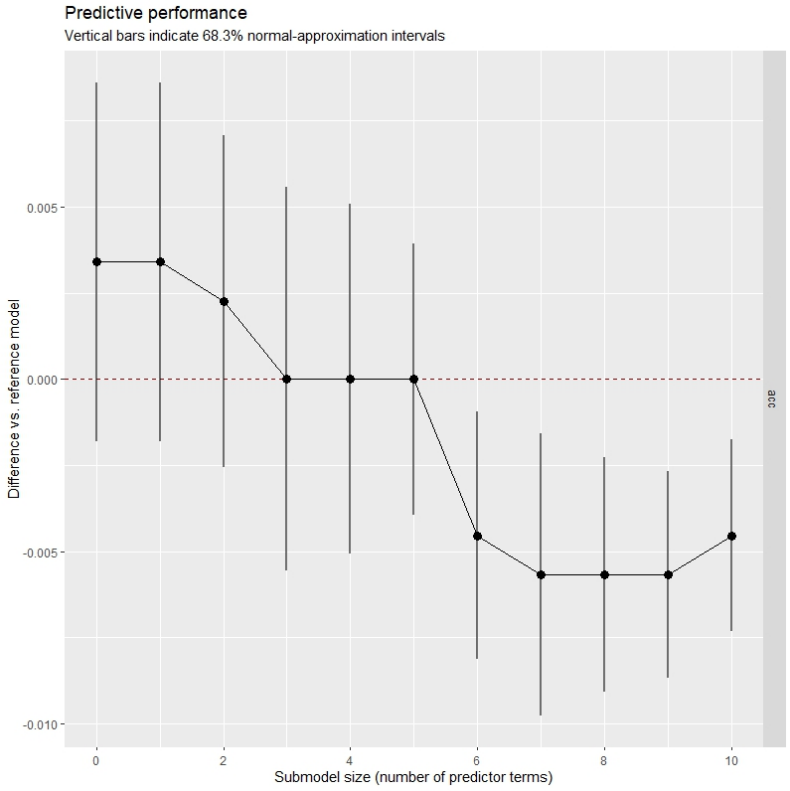


Figure 2.10: Predictive performance in terms of accuracy: difference between each sub-model and the full reference model.

These summaries indicate that a succinct, clinically interpretable subset—dominated by airway procedures and metabolic/respiratory comorbidities—captures most of the predictive content identified by the full model. This result aligns with the predictive, out-of-

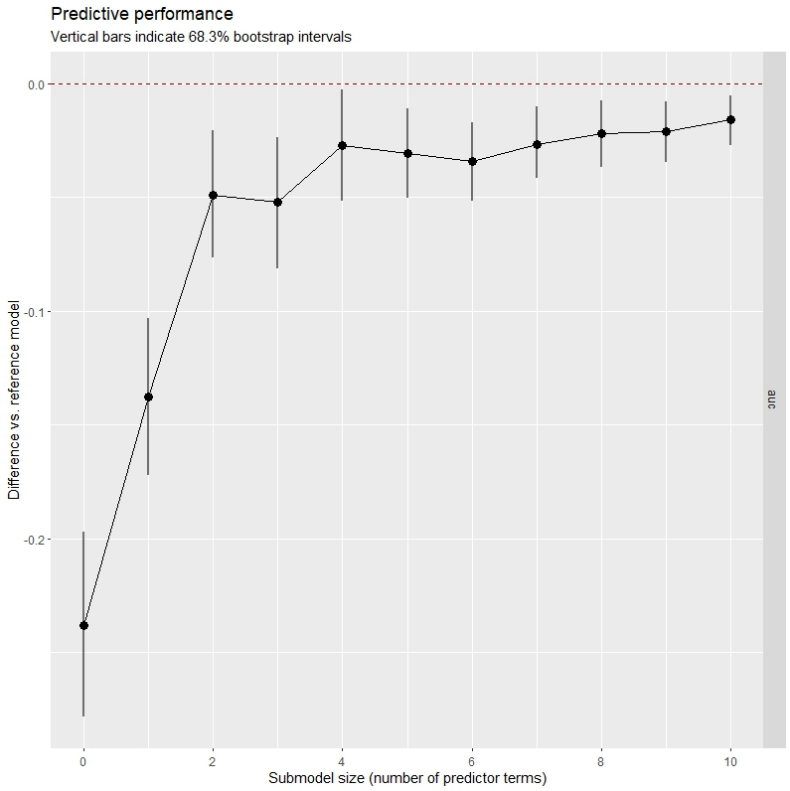


Figure 2.11: Predictive performance in terms of AUC: difference between each sub-model and the full reference model.

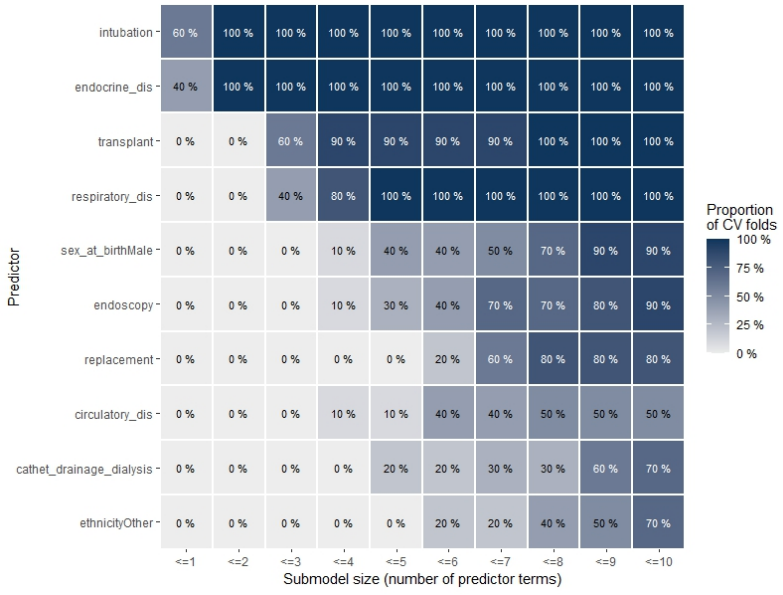


Figure 2.12: Variable inclusion frequency across sub-models of increasing size (10-fold CV).

sample perspective established earlier and complements the stepwise Frequentist screen by providing an explicitly reference-model–based path to simplification with formal cross-validated diagnostics [46, 49, 63].

2.5 Machine Learning Models for Probabilistic Inference and Prediction

Building on shrinkage-based estimation and principled model selection in the previous section, we now evaluate machine–learning models for probabilistic inference and prediction using the same All of Us cohort and binary endpoint (infection within 180 days after the index procedure). The objective is to preserve clinical interpretability while improving discrimination, calibration, and robustness across heterogeneous covariates drawn from electronic records and procedure histories. To keep results comparable with earlier GLM analyses, all models are trained and assessed within a common R workflow, predictions are expressed as probabilities on the event scale, and performance is summarized with threshold–free and thresholded metrics (ROC/AUC, calibration, and decision thresholds) reported alongside class balancing and explicit threshold adjustment when needed [55, 63].

Three complementary families are considered. First, Naive Bayes is estimated under both maximum–likelihood and Bayesian priors to contrast purely empirical frequency estimates with prior–regularized probability models; this pairing clarifies how prior information influences calibration and class separation without altering the basic conditional–independence structure [59]. Second, Bayesian networks encode conditional dependencies among predictors and the outcome, allowing both clinically specified structure (a domain–informed DAG) and data–driven alternatives to be learned and compared. Structure learning is carried out with whitelist/blacklist constraints and, for hybrid discovery, 1000 bootstrap replicates with an arc–strength threshold of 0.80; parameters are then estimated with `bn.fit`, using an imaginary–sample–size (ISS) setting of 88 for the Clinical, Hill–Climbing, and Tabu DAGs, and ISS of 20 for the Bootstrap–averaged Hybrid, so that uncertainty is propagated coherently in the discrete–probabilistic framework [55]. Third, Bayesian Additive Regression Trees (BART) provide a nonparametric ensemble that captures nonlinearity and interactions while retaining probabilistic predictions; interpretation is supported by calibration diagnostics and split–count variable–importance to relate predictive gains back to clinical constructs [46].

Evaluation proceeds in a parallel manner across model families to maintain harmony with the earlier chapters: distribution of predicted probabilities by true class, calibration curves, and ROC comparisons establish probability quality and discrimination; for Bayesian networks, both the learned structures and their operating characteristics are presented to link qualitative graph features with quantitative performance; for BART, calibration and variable–importance connect black–box accuracy to clinical plausibility. Throughout, threshold adjustment is reported where class balance or operating–point considerations are salient, and uncertainty is summarized with cross–validated estimates to emphasize out–of–sample behavior rather than in–sample fit [46, 63]. These models provide a coherent extension from regression to probabilistic machine learning, with design choices (priors, structure constraints, bootstrap aggregation, and ensemble tuning) made explicit so that improvements in prediction can be reconciled with the interpretive needs

of clinical decision–making in a real–world cohort.

2.5.1 Naive Bayes Classifier

Working with the same cohort and endpoint, we trained Naive Bayes classifiers under two estimation regimes—maximum likelihood (Laplace = 0) and Bayesian with Laplace smoothing (pseudo–count = 1)—using a 75%/25% train–test split and uniform class priors (0.5, 0.5) to keep comparisons transparent across models. The test–set results in Table 2.13 summarize accuracy, precision, recall, F1, and Brier score, while Figures 2.13–2.15 provide probability distributions by true class, calibration, and ROC curves.

Table 2.13: Performance comparison between MLE and Bayesian Naive Bayes classifiers on the test dataset

Metric	**MLE (Laplace = 0)**	**Bayesian (Laplace = 1)**
Accuracy	0.765	0.783
Precision	0.940	0.941
Recall	0.788	0.808
F1-score	0.857	0.870
Brier Score	0.445	0.457

The Bayesian model attained slightly higher accuracy (0.783) and recall (0.808) than the MLE variant (0.765 and 0.788, respectively), yielding a higher F1 (0.870 vs. 0.857) driven by improved sensitivity; by contrast, the MLE model achieved a lower Brier score (0.445 vs. 0.457), indicating marginally sharper probability calibration in this dataset. Together these metrics articulate a practical trade–off: smoothing improves case detection at a small calibration cost, a pattern consistent with the role of priors in moderating extreme likelihood–based probabilities in sparse strata [55, 59].

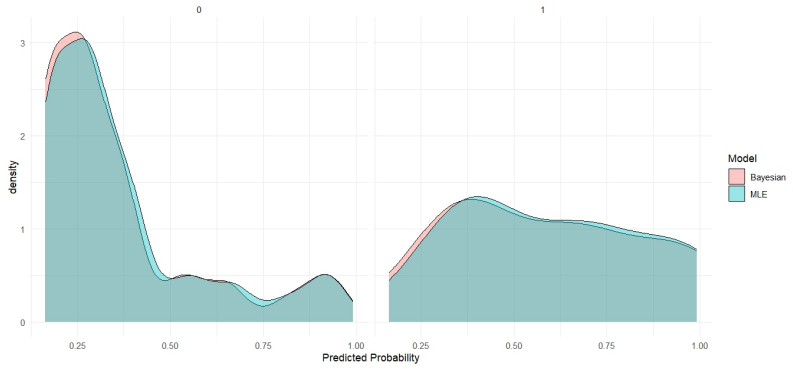


Figure 2.13: Distribution of predicted probabilities by model and true class.

The distributional diagnostics in Figure 2.13 reflect this mechanism directly: MLE predictions are more extreme—especially for non–infected cases—whereas Bayesian smoothing tempers tails and reduces overconfident assignments, bringing probability mass inward. Calibration for the Bayesian model (Figure 2.14) follows the identity line reasonably well

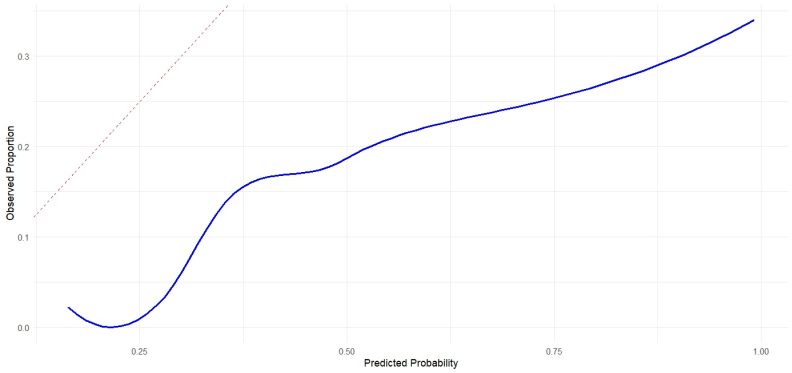


Figure 2.14: Calibration plot for Bayesian Naive Bayes classifier.

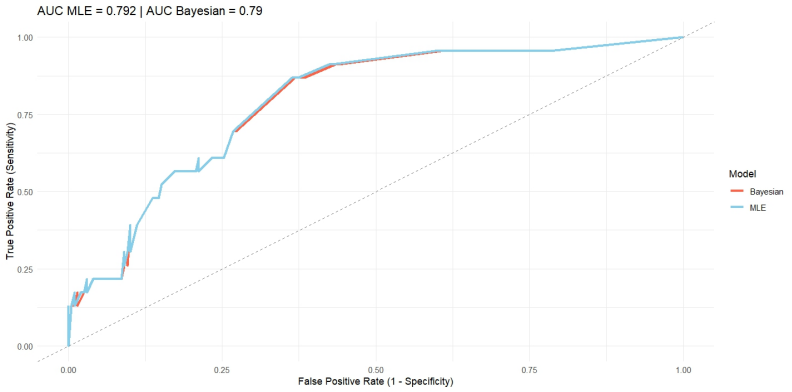


Figure 2.15: ROC curves comparing MLE and Bayesian Naive Bayes models.

with mild underestimation at lower–to–mid risk bands, suggesting that post–hoc calibration (e.g., isotonic or Platt scaling) could further refine risk scores if deployment requires precise absolute probabilities. Discrimination was nearly identical across estimators (AUC 0.792 MLE; 0.790 Bayesian; Figure 2.15), indicating that both approaches rank cases similarly even as their probability sharpness differs; this coherence in ranking, alongside the metric contrasts in Table 2.13, supports reading Naive Bayes as a probability model whose operating characteristics can be tuned via smoothing to align with clinical priorities (recall versus calibration) while remaining computationally simple in R.

2.5.2 Bayesian Network Model

Using the same cohort and endpoint, Bayesian networks were estimated to capture conditional dependencies among predictors and the infection outcome, with four complementary structures evaluated for both interpretability and predictive quality. A clinically specified directed acyclic graph (Clinical DAG) encoded domain constraints via whitelist/blacklist rules; two data–driven alternatives were learned with Hill–Climbing (HC) and Tabu search; and a Hybrid model averaged structures obtained from 1,000 bootstrap replicates with an arc–strength threshold of 0.80 to retain only stable relations. Discrete conditional probability tables were then fit with `bn.fit`, using an imaginary–sample–size (ISS) of 88 for the Clinical, HC, and Tabu DAGs and ISS of 20 for the Bootstrap–averaged Hybrid so that parameter uncertainty was coherently regularized within each graph. Class balancing and threshold adjustment were applied at the scoring stage to maintain comparability across models on the same test split, and reporting emphasizes accuracy, precision, recall, and F1 alongside ROC curves to connect graph structure with operating characteristics [55, 59].

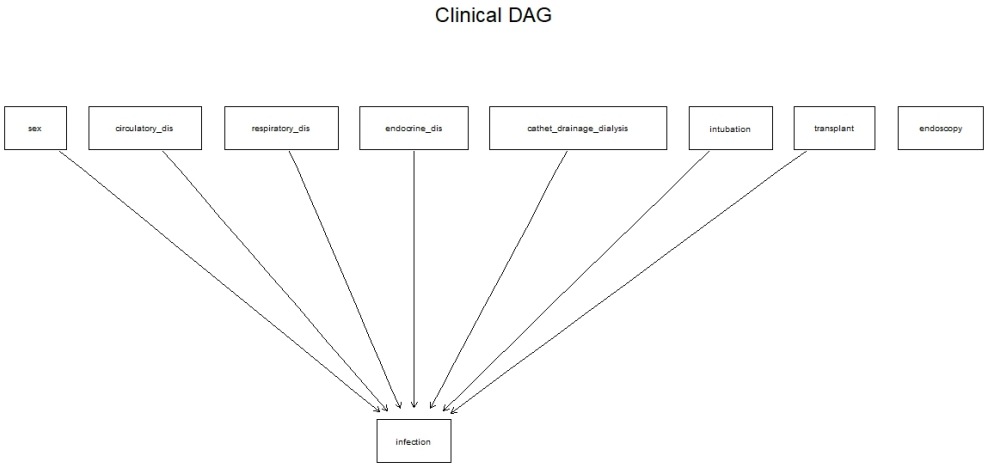


Figure 2.16: Clinical DAG.

The four learned structures are summarized in Figures 2.16–2.19. The Clinical DAG foregrounds links consistent with prior evidence—procedural exposures and cardiopulmonary/endocrine comorbidity pointing toward infection—while suppressing implausible back–doors. The HC and Tabu graphs recover similar backbones but differ in a few secondary arcs, reflecting their distinct search heuristics and regularization trade–offs. The

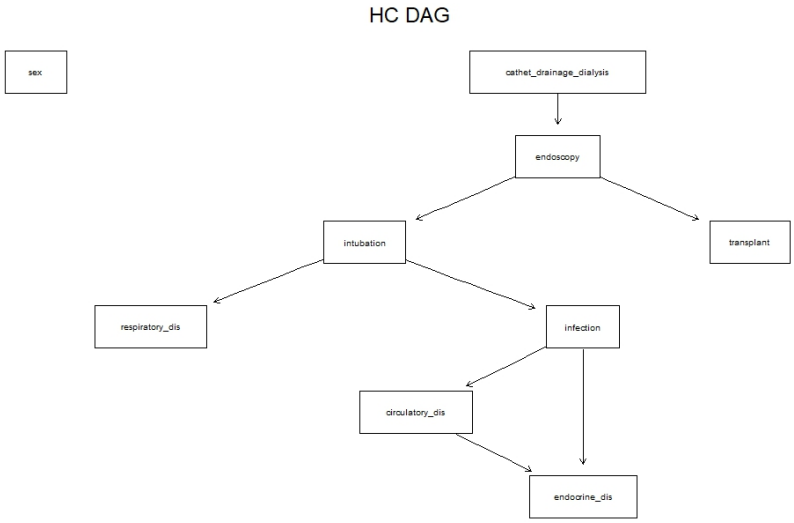


Figure 2.17: Hill-Climbing DAG.

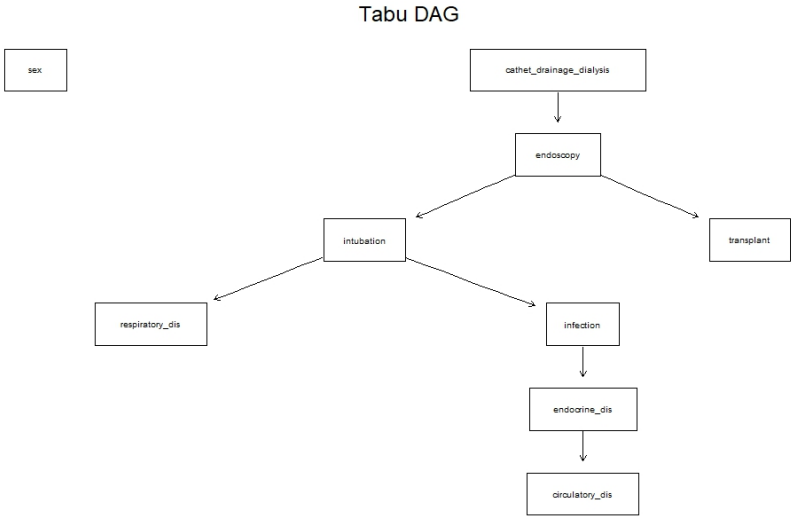


Figure 2.18: Tabu DAG.

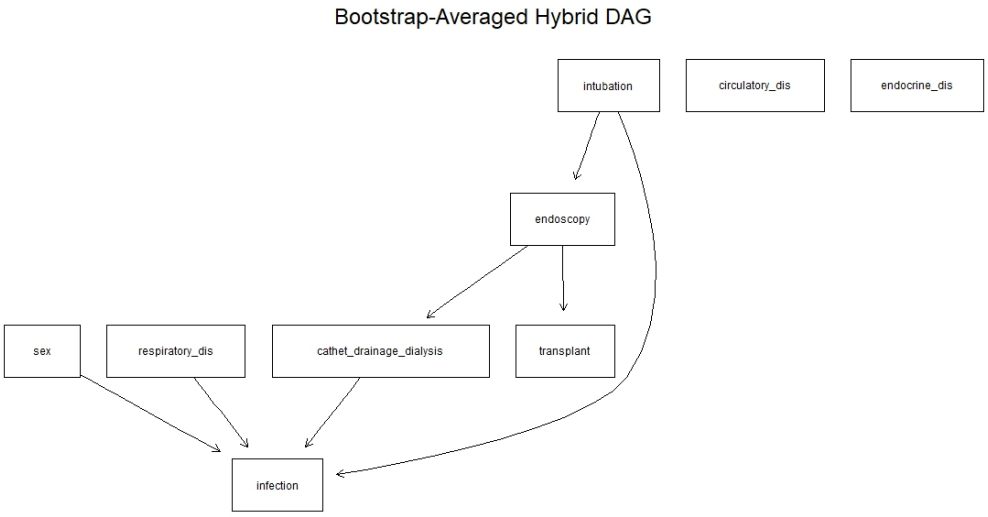


Figure 2.19: Bootstrap-Averaged Hybrid DAG.

Hybrid Bootstrap DAG is visually sparser yet retains high–strength relations: arcs from intubation and endocrine disorder into the outcome appear consistently, with respiratory disease and transplant also recurrent; edges lacking bootstrap support are pruned, improving stability. This qualitative harmony across structures anticipates the quantitative ranking.

Table 2.14: Comparison of Model Performance across Bayesian Network Approaches (Accuracy, Precision, Recall, F1)

Model	**Accuracy**	**Precision**	**Recall**	**F1**
Clinical DAG	0.859	0.915	0.929	0.922
Hill-Climbing DAG	0.826	0.908	0.894	0.901
Tabu DAG	0.792	0.937	0.823	0.876
Hybrid (Bootstrap-Averaged)	0.896	0.896	1.000	0.945

Performance results on the held–out data (Table 2.14) show that the Bootstrap–averaged Hybrid provides the best overall operating point, achieving the highest accuracy with perfect recall and the strongest F1 among the four approaches (Accuracy = 0.8959, Recall = 1.0000, F1 = 0.9451). The Clinical DAG follows closely across metrics, offering competitive discrimination with strong clinical interpretability. HC and Tabu deliver similar accuracy but slightly lower recall/F1, consistent with their denser yet less stable arc sets.

ROC curves in Figures 2.20–2.23 corroborate this ranking: the Hybrid dominates across thresholds, the Clinical DAG tracks with marginally reduced sensitivity at intermediate operating points, and HC/Tabu curves lie just below. In practical terms, the bootstrap aggregation concentrates graph mass on clinically coherent predictors—intubation, endocrine and respiratory disorders, and transplant—so that threshold adjustments can prioritize recall (case finding) without sacrificing precision unduly.

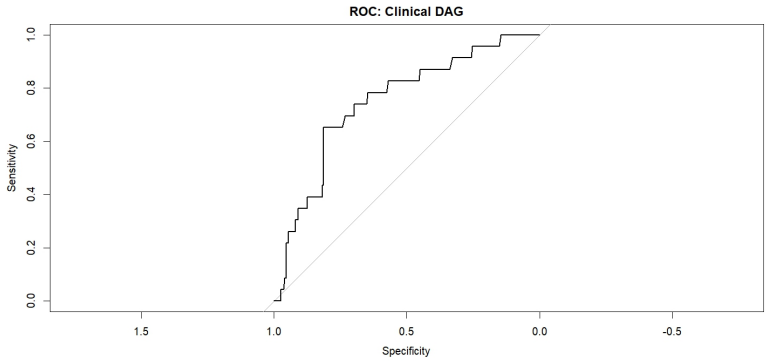


Figure 2.20: Clinical DAG ROC.

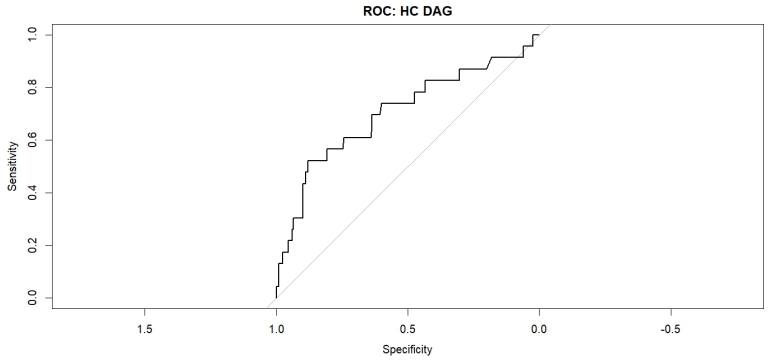


Figure 2.21: Hill-Climbing DAG ROC.

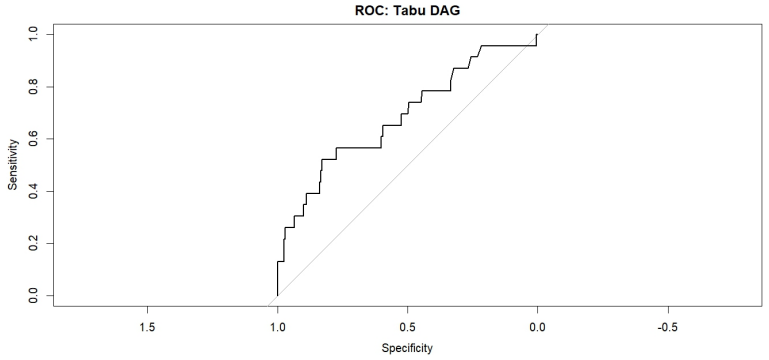


Figure 2.22: Tabu DAG ROC.

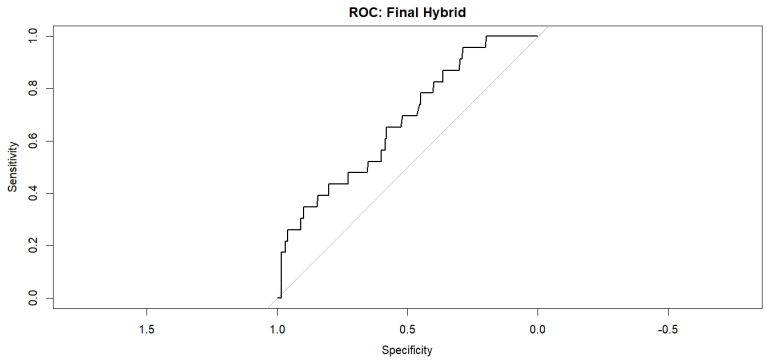


Figure 2.23: Bootstrap-Averaged Hybrid DAG ROC.

2.5.3 Bayesian Additive Regression Trees (BART)

Using the same cohort and endpoint, BART was fitted for probabilistic classification with predictions evaluated on a held-out test set. A data-driven operating point was selected by maximizing Youden's index from the ROC curve, yielding threshold performance summarized in Table 2.15; discrimination and probability quality are presented in Figures 2.24–2.25, and model interpretability is supported by split-count variable importance in Figure 2.26.

Table 2.15: Performance Metrics for BART Model Using Optimal Threshold

Metric	**Value**
Accuracy	0.661
Kappa	0.204
Balanced Accuracy	0.734
Sensitivity (Recall)	0.826
Specificity	0.641
Positive Predictive Value (Precision)	0.211
Negative Predictive Value	0.969
Detection Rate	0.086

On the test set, sensitivity was high (0.826) with a correspondingly strong negative predictive value (0.969), while specificity was more modest (0.641), producing an overall balanced accuracy of 0.734 and an accuracy of 0.661 at the chosen operating point; this pattern reflects an emphasis on case-finding with tolerable false positives in the clinical context. The ROC curve (Figure 2.24) indicates good discrimination (AUC = 0.78), consistent with effective nonlinear risk stratification, whereas the calibration plot (Figure 2.25) shows near-identity alignment overall with mild underestimation at lower predicted risks; in settings where absolute risk is critical, post-hoc calibration could further refine probabilities while leaving ranking intact [55, 63].

Variable-importance diagnostics (Figure 2.26) highlight endocrine disorder, intubation, and transplant as the most influential predictors, with respiratory disease also contributing, a profile that coheres with prior GLM and Bayesian-network results and supports clinical face validity of the learned signal.

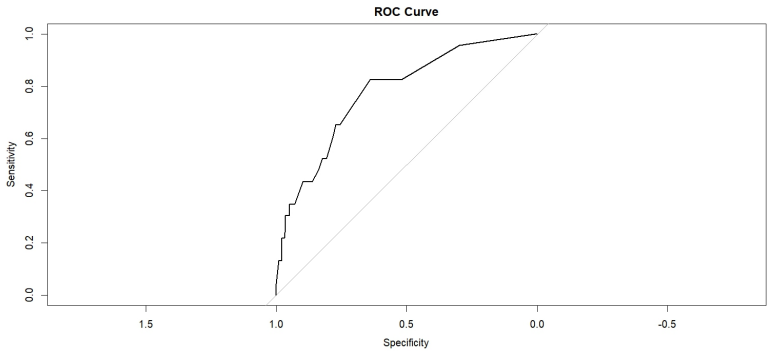


Figure 2.24: ROC of BART.

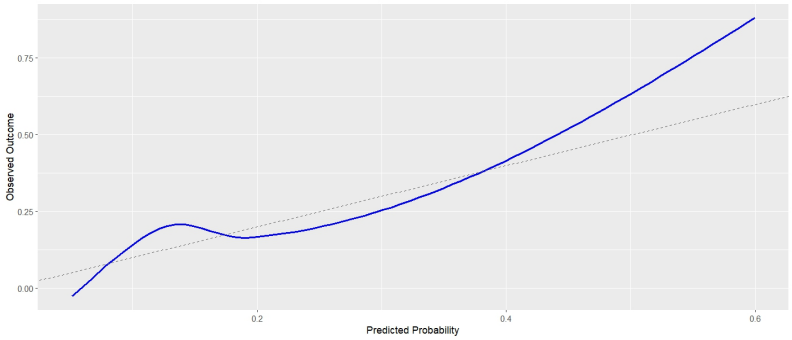


Figure 2.25: Calibration Plot of BART.

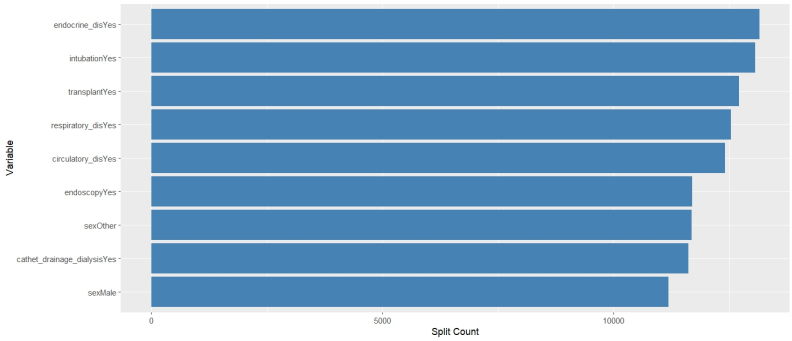


Figure 2.26: Variable Importance (Split Counts).

2.6 Generalized Linear Models in Infectious Disease Analysis and Surveillance: Methods for Correlated Data

Building on shrinkage–based estimation, model selection, and probabilistic prediction, the present section addresses outcomes that arise within clustered or longitudinal structures, where observations from the same patient or clinical setting are correlated. In such settings, treating records as independent can bias uncertainty estimates and distort effect sizes, motivating models that explicitly represent multi–level variation and within–cluster dependence. GLMM extend GLM by introducing random effects to capture subject– or cluster–specific departures from the population mean, thereby preserving the familiar link–function framework while accounting for correlation in the data [47]. Random–intercept and random–slope terms allow baseline risk and time–varying trends to vary across patients or wards, and the corresponding intraclass correlation coefficient (ICC) provides a summary of within–cluster similarity that is interpretable alongside fixed–effect estimates.

From a likelihood perspective, GLMM parameters are estimated by maximizing a marginal likelihood obtained by integrating over the random effects. Approximations such as Laplace and adaptive Gaussian quadrature enable stable computation in logistic and count models, yielding effect estimates and standard errors that align with epidemiologic reporting on the odds–ratio or rate–ratio scales [48]. Diagnostics remain central: scaled residual summaries, variance–component estimates, and random–effect correlations are read in concert to evaluate fit and to detect misspecification that might masquerade as a fixed–effect signal.

A complementary Bayesian formulation treats random effects and variance components as parameters with prior distributions, inducing partial pooling across clusters and propagating uncertainty coherently through all levels of the hierarchy. Hyperpriors on variance components regularize extreme estimates in sparse strata, while posterior distributions for fixed effects, random effects, and ICC quantify uncertainty in a unified way [64]. Computation relies on gradient–based Markov chain Monte Carlo (HMC with the No–U–Turn Sampler), with convergence assessed via rank–normalized $\hat{R}$ and effective sample sizes; posterior predictive checks support model adequacy by aligning fitted and observed cluster patterns [46, 49, 50].

To maintain continuity with earlier sections, analyses are conducted on the same All of Us cohort and binary endpoint (infection within 180 days after the index procedure), and results are reported on epidemiologically familiar scales. Conceptual illustrations of correlation structures and random–effect behavior are provided for orientation, whereas the main narrative prioritizes empirical results from three GLMM specifications and a set of Bayesian hierarchical models, including posterior distributions, ICC estimates, and random–effect diagnostics. The hierarchical analyses, alongside the previously developed independent–data models and machine–learning predictors, complete a coherent workflow in which heterogeneity and dependence are modeled explicitly, uncertainty is calibrated for clustered designs, and clinical interpretability is preserved within a unified statistical framework [46, 47].

2.6.1 Generalized Linear Mixed Models

We interpreted three logistic GLMM fitted to the All of Us longitudinal cohort—progressing from a univariable random–intercept model (GLMM 1) through a multivariable random–intercept model (GLMM 2) to a random–intercept–plus–random–slope specification for *intubation* (GLMM 3)—to quantify subject–level heterogeneity and covariates effects on infection within 180 days after the index procedure. Model comparison used likelihood-based criteria (AIC, BIC, log-likelihood, deviance), with random–effect variance components and scaled residual summaries guiding adequacy checks, in line with standard GLMM practice [47, 48].

Table 2.16: Model fit statistics (GLMM 1)

Model Fit Statistic	**Value**
AIC	3753.8
BIC	3773.8
Log-Likelihood	−1873.9
Deviance	3747.8
Residual degrees of freedom	5669

Table 2.17: Random effect variances (GLMM 1)

Random Effect	**Variance**	**Std. Dev.**
id (intercept)	2.903	1.704

Table 2.18: Summary of Scaled Residuals (GLMM 1)

Min	**Q1**	**Median**	**Q3**	**Max**
−1.8613	−0.3016	−0.1962	−0.1401	7.6346

Table 2.16 presents the standard model fit statistics, including the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), log-likelihood, and deviance from GLMM 1. The age-only specification with a subject-level random intercept establishes a baseline for within–person clustering. Fit indices () indicate a reference level of explanatory power for subsequent comparisons, while the estimated random–intercept variance (2.903; SD 1.704) signals substantial between–subject heterogeneity (Table 2.17). Scaled residuals show mild left skew but no extreme deviation (Table 2.18). The fixed effect for age $\geq$ 60 is positive and precise ($\hat{\beta} = 0.607$, $p < 0.001$), corresponding to an odds ratio of $\exp(0.607) \approx 1.84$ for older versus younger participants (Table 2.19). The intraclass correlation (ICC) is approximately 46%, underscoring the importance of accounting for within–subject dependence in this setting, which is consistent with the interpretive framework for subject-specific versus marginal effects in mixed models [47]. These results provide the subject–level baseline against which richer specifications are read.

Table 2.19: Fixed effects estimates (GLMM 1)

Fixed Effect	**Estimate**	**Std. Error**	**z value**	**Pr(>** $\lvert z \rvert$**)**
Intercept	−3.1491	0.1620	−19.436	< 0.001
Age ≥ 60	0.6072	0.1343	4.521	< 0.001

Table 2.20: Model fit statistics (GLMM 2)

Model Fit Statistic	**Value**
AIC	3532.7
BIC	3605.8
Log-Likelihood	−1755.3
Deviance	3510.7
Residual degrees of freedom	5661

Table 2.21: Random effect variances (GLMM 2)

Random Effect	**Variance**	**Std. Dev.**
id (intercept)	3.475	1.864

Table 2.22: Summary of Scaled Residuals (GLMM 2)

Min	**Q1**	**Median**	**Q3**	**Max**
−2.7489	−0.2965	−0.1717	−0.0994	7.5862

Table 2.23: Fixed effects estimates (GLMM 2)

Fixed Effect	**Estimate**	**Std. Error**	**z value**	**Pr(>** $\lvert z \rvert$**)**
Intercept	−4.0663	0.2519	−16.145	< 0.001
Sex at birth: Male	0.1252	0.2135	0.586	0.558
Race: White	−0.2314	0.2183	−1.060	0.289
Ethnicity: Other	−0.0808	0.6342	−0.127	0.899
Respiratory disorder	1.7029	0.3707	4.594	< 0.001
Endocrine disorder	1.6020	0.1830	8.756	< 0.001
Cancer	1.3946	0.1659	8.409	< 0.001
Catheter/Dialysis/Drainage	0.6909	0.2525	2.736	0.006
Intubation	1.4369	0.3136	4.582	< 0.001
Transplant	−0.0754	0.2473	−0.305	0.760

Compared to the univariable model, the multivariable specification shows improved model fit (Table 2.20). Adding the clinically motivated covariates improves fit (lower AIC/BIC and deviance; Table 2.20) while retaining a sizable random–intercept variance (3.475; SD 1.864; Table 2.21), indicating persistent baseline heterogeneity even after adjustment. Residual summaries remain stable (Table 2.22). Fixed-effect estimates (Table 2.23) highlight strong positive associations for respiratory, endocrine, and cancer comorbidities, and for intubation; transplant is imprecisely estimated (near null), and demographic indicators contribute little once clinical factors are included. The interpretation remains on the odds–ratio scale familiar in epidemiology, while the random effect absorbs unexplained subject–level risk.

Table 2.24: Model fit statistics (GLMM 3)

Model Fit Statistic	**Value**
AIC	3532.1
BIC	3618.5
Log-Likelihood	−1753.1
Deviance	3506.1
Residual degrees of freedom	5659

Table 2.25: Random effect variances and correlation (GLMM 3)

Random Effect Term	**Variance**	**Std. Dev.**
id Intercept	3.647	1.910
Intubation (slope)	3.157	1.777
Correlation (Intercept, Intubation)	−0.18	

Table 2.26: Summary of Scaled Residuals (GLMM 3)

Min	**Q1**	**Median**	**Q3**	**Max**
−2.8074	−0.2923	−0.1670	−0.0966	7.7640

Allowing the *intubation* effect to vary by subject (random slope) yields a modest additional improvement in fit relative to GLMM 2 (Table 2.24). Variance components for both the intercept and the *intubation* slope are non–trivial, and their estimated correlation is slightly negative (−0.18), suggesting that higher baseline risk is weakly associated with a smaller incremental *intubation* effect (Table 2.25). Residual summaries are similar to GLMM 2 (Table 2.26). Inferences for the main clinical signals (respiratory, endocrine, cancer) remain directionally stable and significant, while the *intubation* effect persists but with a larger standard error reflecting subject–specific variation. This pattern fits the mixed–model rationale: partial pooling captures heterogeneity without sacrificing interpretability, with likelihood criteria guiding the acceptable complexity of the random–effects structure [47].

2.6.2 Generalized Estimating Equations

To complement the subject–specific inferences from GLMMs, we fitted multivariable logistic GEEs that target population–averaged effects under alternative working correlation structures: independence (GEE 1), exchangeable (GEE 2), and AR(1) (GEE 3). Estimation used `geem()` from `geeM` with robust (sandwich) variance to guard against misspecification of the within–subject correlation, and model comparison relied on QIC rather than likelihood–based criteria, in keeping with quasi–likelihood estimation [65–67]. Coefficient patterns are remarkably coherent with the mixed–model results but are considered here as marginal effects: across specifications, endocrine disorder, cancer, and intubation remained strong positive predictors, respiratory disorder was clearly positive under exchangeable and AR(1) assumptions and borderline under independence, while transplant showed sensitivity to the working correlation (significant in GEE 1 and GEE 3, not in GEE 2). Demographic terms (sex at birth, race, ethnicity) did not show stable associations after clinical adjustment. These contrasts illustrate how the working correlation influences standard errors—and, in turn, p–values—without overturning the substantive ranking of clinical predictors [66, 67]. The QIC ranking favored AR(1), consistent with longitudinal dependence over time, and a side–by–side comparison of robust versus model–based (non–robust) standard errors in GEE 3 underscored the inferential impact of using sandwich corrections in clustered designs [68].

Table 2.27: Estimated Coefficients of GEEs

	GEE 1 (Indep.)		**GEE 2 (Exch.)**		**GEE 3 (AR(1))**	
Variable	β	**Robust SE**	β	**Robust SE**	β	**Robust SE**
(Intercept)	-2.33^{***}	0.26	-2.69^{***}	0.14	-2.27^{***}	0.20
Sex at Birth: Male	-0.33	0.27	-0.01	0.17	-0.21	0.22
Race: White	$-0.65^{\cdot}$	0.33	-0.18	0.17	$-0.40^{\cdot}$	0.23
Ethnicity: Other	$-0.65^{\cdot}$	0.34	0.11	0.38	-0.35	0.36
Respiratory disorder	$0.79^{\cdot}$	0.42	1.15^{***}	0.30	0.99^{**}	0.31
Endocrine disorder	0.68^{*}	0.29	1.05^{***}	0.15	0.74^{***}	0.22
Cancer	1.01^{***}	0.28	1.07^{***}	0.21	0.87^{***}	0.19
Catheter/Dialysis/Drainage	0.34	0.43	0.53	0.35	0.28	0.21
Intubation	1.35^{***}	0.31	1.03^{***}	0.28	0.63^{*}	0.32
Transplant	1.02^{*}	0.40	0.22	0.48	0.53^{*}	0.25

Significance: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.10$.

Table 2.27 reports estimates and robust standard errors under each working correlation. Endocrine disorder, cancer, and intubation are consistently positive with conventional significance; respiratory disorder strengthens when correlation is modeled (GEE 2, GEE 3), reflecting improved efficiency under a plausible dependence structure; transplant's significance varies with the assumed correlation, highlighting a practical modeling choice in longitudinal surveillance analyses.

Table 2.28 compares models using QIC and quasi–likelihood; AR(1) yields the low-

Table 2.28: Model Fit for GEEs

Model	QIC	Quasi-likelihood	$\hat{\phi}$	$\hat{\alpha}$	Correlation Structure
GEE 1	4201.84	−2011.07	1.0200	0.0000	Independence
GEE 2	4170.70	−2072.69	0.9379	0.2330	Exchangeable
GEE 3	**4124.30**	−2033.75	0.8835	**0.5869**	AR(1)

est QIC and the largest estimated within–subject correlation parameter $\hat{\alpha}$, supporting time–ordered dependence of repeated observations in this cohort.

Table 2.29: Comparison of Robust vs Non-Robust Standard Errors (GEE 3)

Variable	Robust SE		Non-Robust SE	
	OR	95% CI	OR	95% CI
(Intercept)	0.10	[0.07, 0.15]	0.10	[0.08, 0.13]
Sex at Birth: Male	0.81	[0.53, 1.25]	0.81	[0.63, 1.04]
Race: White	0.67	[0.42, 1.06]	0.67	[0.51, 0.88]
Ethnicity: Other	0.71	[0.35, 1.42]	0.71	[0.32, 1.54]
Respiratory Disorder	2.69	[1.47, 4.93]	2.69	[1.77, 4.10]
Endocrine Disorder	2.10	[1.38, 3.21]	2.10	[1.68, 2.63]
Cancer	2.40	[1.64, 3.50]	2.40	[1.90, 3.02]
Catheter/Dialysis/Drainage	1.32	[0.87, 2.00]	1.32	[0.97, 1.81]
Intubation	1.87	[1.01, 3.48]	1.87	[1.29, 2.73]
Transplant	1.69	[1.04, 2.76]	1.69	[1.32, 2.17]

Finally, Table 2.29 contrasts inference for two clinically salient predictors under GEE 3 when using robust (sandwich) versus conventional model–based standard errors. Odds–ratio point estimates are identical by construction, yet interval widths differ: for *intubation*, the non–robust CI excludes 1 decisively, while the robust CI is wider and only marginally excludes 1; for *transplant*, robust intervals are likewise wider and more conservative. This illustrates why sandwich corrections are recommended when the true correlation is uncertain or misspecified [67, 68].

2.6.3 Bayesian Hierarchical Models

We fitted four Bayesian hierarchical logistic models (denoted BHM 1 - BHM 4) to the same longitudinal cohort analyzed in the preceding sections, progressing from a sparse random–intercept specification to a richer structure with random slopes and finally to an informative–prior formulation. Computation proceeded in brms/Stan with four chains, and convergence was assessed via rank–normalized $\hat{R}$ and effective sample sizes; posterior summaries (means and 95% credible intervals) are reported alongside intraclass correlation (ICC) where appropriate [36, 46, 49, 50]. In line with the GLMM results, inference emphasizes clinically interpretable log–odds effects while explicitly propagating between–subject heterogeneity.

Table 2.30: BHM 1. Posterior Estimates

Parameter	**Estimate**	**SE**	**2.5%**	**97.5%**	$\hat{R}$
Multilevel Hyperparameters					
sd(Intercept)	1.63	0.13	1.40	1.90	1.01
Regression Coefficients					
Intercept	−3.05	0.15	−3.35	−2.76	1.00
Age ≥ 60	0.60	0.13	0.34	0.87	1.00

Table 2.31: BHM 2. Posterior Estimates

Parameter	**Estimate**	**SE**	**2.5%**	**97.5%**	$\hat{R}$
Multilevel Hyperparameters					
sd(Intercept)	1.64	0.13	1.40	1.90	1.01
Regression Coefficients					
Intercept	−3.04	0.14	−3.31	−2.78	1.00
Age ≥ 60 (informative prior)	0.57	0.09	0.40	0.75	1.00

BHM 1 and BHM 2 include a subject–level random intercept and a single fixed effect for age ≥ 60; the sole difference is the prior on the age coefficient, which is diffuse in BHM 1 and literature–informed in BHM 2 (Normal mean 0.554, SD 0.118 on the log–odds scale) — reflecting meta–analytic evidence for elevated risk in older adults. Across both models, the random–intercept SD is stable (about 1.63–1.64), indicating substantial between–subject heterogeneity, and the age effect remains positive with narrower intervals under the informative prior (Tables 2.30–2.31).

Figures 2.27–2.28 visualize the posterior for BHM 1 and its ICC: using the logistic–link convention $\sigma_e^2 = \pi^2/3$, the ICC centers near 0.44 (95% CrI 0.37, 0.52), consistent with strong clustering at the subject level.

Figure 2.29 displays the posterior distributions of the subject-specific random intercepts for the first 15 patients. These distributions capture how each individual's baseline risk deviates from the population-level average. Visualizing the heterogeneity in these intercepts can offer insight into latent subject-level variability and helps identify systematic patterns associated with increased or decreased risk.

Extending the linear predictor to allow individual–specific departures for *intubation* sharpened our understanding of heterogeneity. Table 2.32 reports the full posterior summary. Posterior SDs indicate sizeable baseline dispersion (intercept SD 1.83, CrI $[1.53, 2.17]$) and uncertain slope variation (intubation SD 2.19, CrI $[0.45, 4.20]$), with weak and imprecise correlation between intercept and slope (0.05, CrI $[-0.58, 0.74]$). Fixed–effect patterns mirror the GLMM: respiratory, endocrine, and cancer disorders are robustly positive, while demographic terms remain close to null; the population–level *intubation* effect is positive but uncertain (mean 1.03, CrI $[-0.23, 1.95]$).

Figure 2.30 shows that subject–specific random slopes concentrate near zero, consistent with partial pooling given sparse *intubation* exposure. We then incorporated evidence–based priors for selected covariates (sex, endocrine disorder, cancer, catheter/-dialysis/drainage, and *intubation*), summarized in Table 2.33.

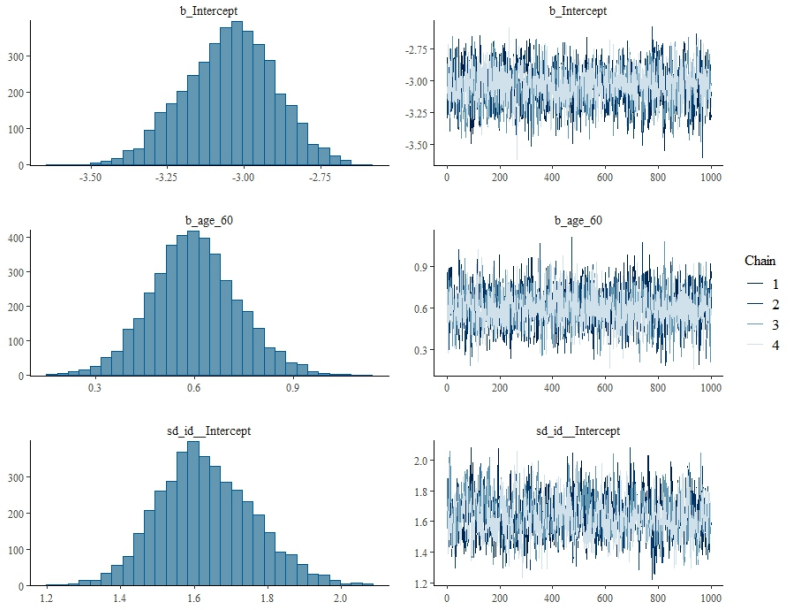


Figure 2.27: BHM 1. Posterior distribution plots.

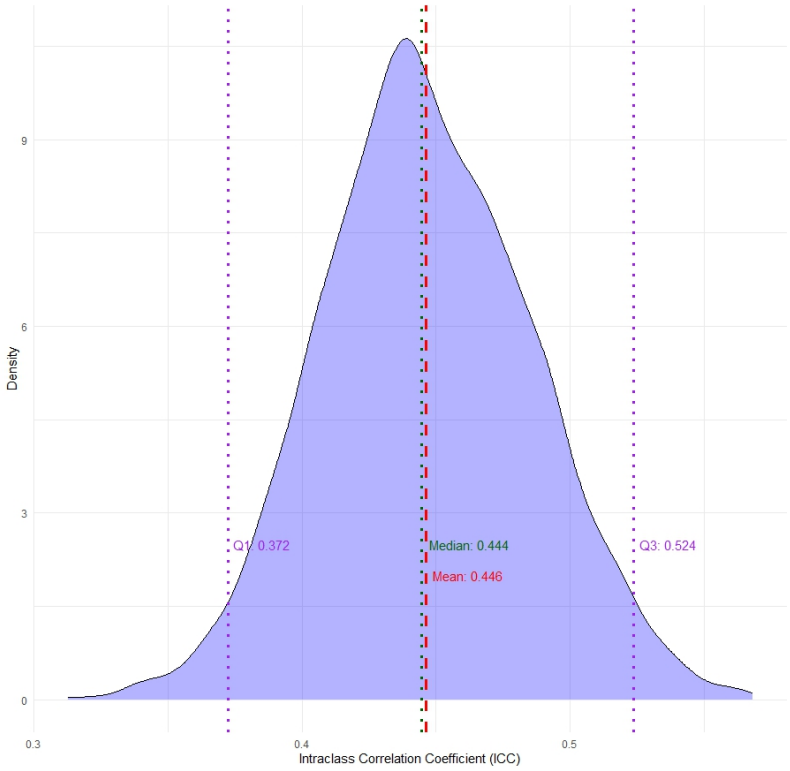


Figure 2.28: BHM 1. Intraclass Correlation Coefficient.

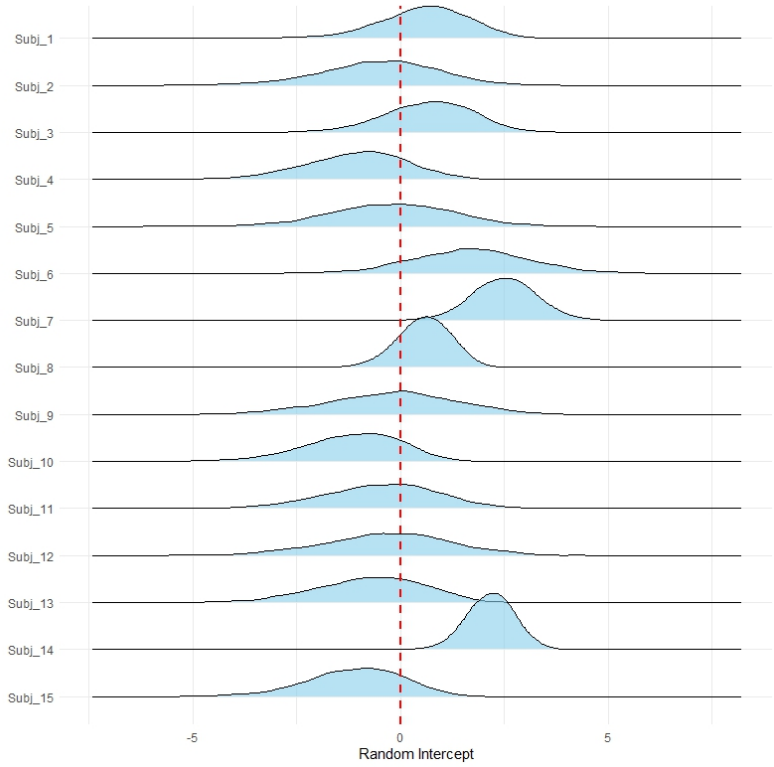


Figure 2.29: BHM 1. Random effects on intercept (subjects 1–15).

Table 2.32: Posterior Estimates for BHM 3

Parameter	**Estimate**	**SE**	**2.5%**	**97.5%**	$\hat{R}$
Multilevel Hyperparameters					
sd(Intercept)	1.83	0.16	1.53	2.17	1.00
sd(Intubation)	2.19	0.92	0.45	4.20	1.01
cor(Intercept, Intubation)	0.05	0.34	−0.58	0.74	1.00
Regression Coefficients					
Intercept	−4.04	0.25	−4.56	−3.58	1.00
Sex: Male	0.14	0.21	−0.28	0.56	1.00
Race: White	−0.24	0.22	−0.67	0.19	1.00
Ethnicity: Other	−0.11	0.63	−1.35	1.08	1.00
Respiratory Disorder	1.64	0.37	0.92	2.34	1.00
Endocrine Disorder	1.65	0.19	1.30	2.04	1.00
Cancer	1.42	0.17	1.10	1.76	1.00
Catheter/Dialysis/Drainage	0.68	0.25	0.17	1.15	1.00
Intubation	1.03	0.55	−0.23	1.95	1.00
Transplant	−0.08	0.25	−0.57	0.40	1.00

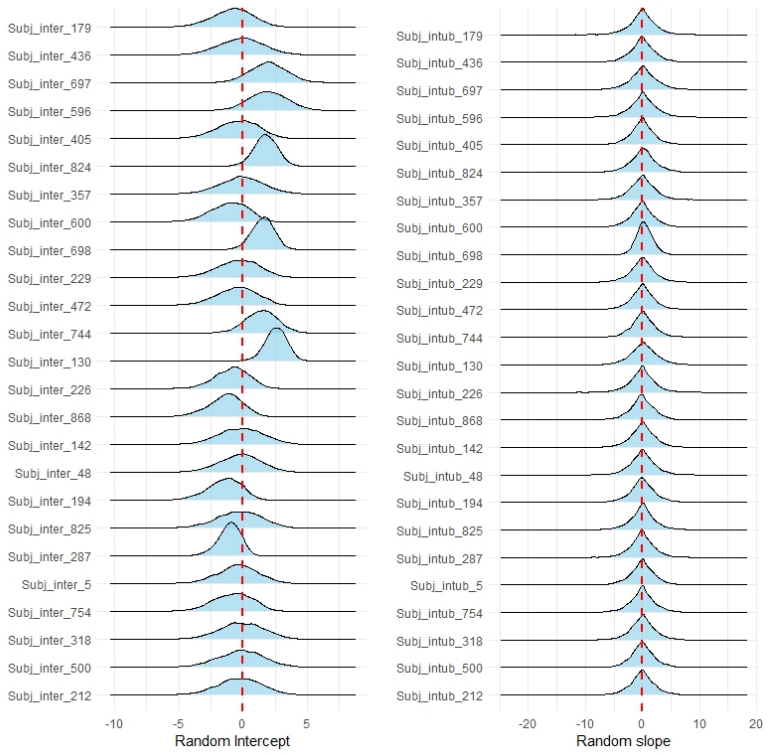


Figure 2.30: BHM 3. Random intercepts and random slopes for *intubation* (sampled subjects).

Table 2.33: Summary of Priors from Literature Review

Variable	**Mean**	**Std. Dev.**
Sex (Male)	0.29	0.05
Endocrine Disorder	0.83	0.32
Cancer	0.67	0.33
Catheter/Dialysis/Drainage	1.04	0.20
Intubation	1.52	0.40

Relative to BHM 3, posterior uncertainty decreased for these targets, and the *intubation* coefficient became more precise and decisively positive (mean 1.36, CrI $[0.77, 1.91]$), aligning with the direction observed under likelihood–based and noninformative Bayesian analyses. Random–effect SDs modestly decreased and the intercept–slope correlation remained weak. Figure 2.31 contrasts marginal posteriors under default versus informative priors, while Table 2.34 details the updated posterior estimates.

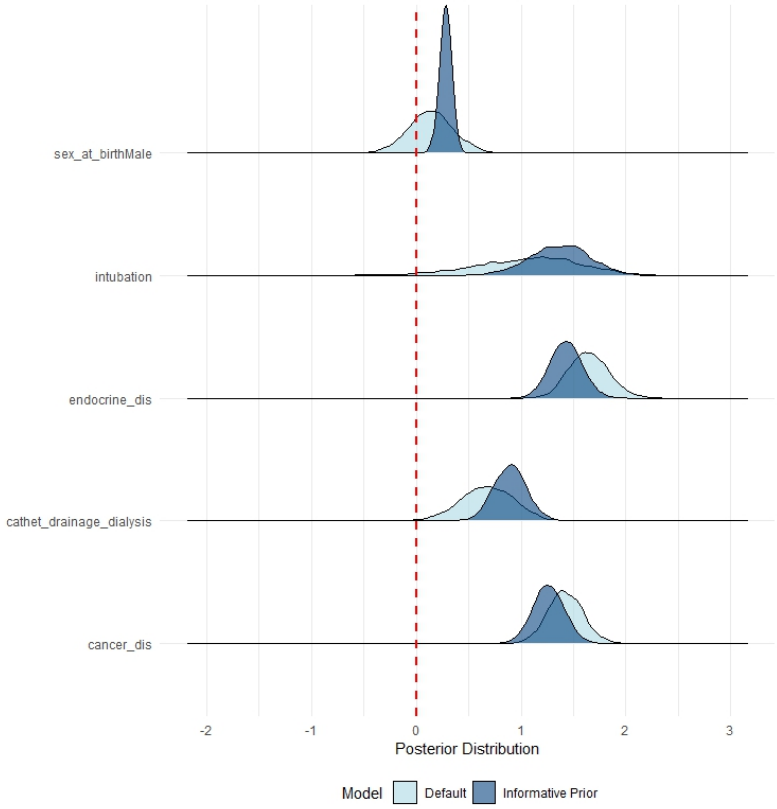


Figure 2.31: BHM 3 vs BHM 4. Posterior distributions (default vs informative priors) for selected coefficients.

Together, the sequence BHM 1→BHM 4 illustrates how hierarchical modeling and prior information complement each other: clustering and shrinkage stabilize subject–specific variation, while external evidence sharpens inference for clinically salient effects without overturning the substantive signal identified earlier.

2.7 Residuals and Overdispersion in Generalized Linear Models

Building on the regression, selection, and hierarchical modeling developed in earlier sections, this section turns to diagnostic assessment and robust modeling for binary and count outcomes when dispersion and zero–inflation challenge standard assumptions. In routine surveillance and clinical cohorts, event processes routinely display variance that exceeds the mean and an excess of structural zeros; relying on canonical likelihoods without

Table 2.34: BHM 4. Posterior Estimates with Informative Priors

Parameter	**Estimate**	**SE**	**2.5%**	**97.5%**	$\hat{R}$
Multilevel Hyperparameters					
sd(Intercept)	1.73	0.14	1.46	2.01	1.01
sd(Intubation)	1.76	0.81	0.24	3.49	1.01
cor(Intercept, Intubation)	0.02	0.35	−0.60	0.76	1.00
Regression Coefficients					
Intercept	−3.91	0.21	−4.33	−3.51	1.00
Sex: Male	0.28	0.05	0.19	0.37	1.00
Race: White	−0.21	0.21	−0.64	0.19	1.00
Ethnicity: Other	−0.08	0.62	−1.32	1.13	1.00
Respiratory Disorder	1.55	0.36	0.86	2.24	1.00
Endocrine Disorder	1.43	0.15	1.14	1.73	1.00
Cancer	1.26	0.15	0.97	1.55	1.00
Catheter/Dialysis/Drainage	0.90	0.16	0.59	1.21	1.00
Intubation	1.36	0.29	0.77	1.91	1.00
Transplant	0.03	0.24	−0.43	0.49	1.00

checking these features can bias uncertainty and distort effect sizes [47]. Accordingly, we organize the analysis around two linked goals: (i) to diagnose misfit on the probability and count scales using residual displays tailored to generalized models, and (ii) to compare models that explicitly accommodate overdispersion and zero–inflation while preserving epidemiologic interpretability.

Diagnostics begin from fitted GLM, where binned residual plots and simulation–based residual checks provide visual evidence for lack of calibration across the range of fitted values and for patterns consistent with dispersion beyond the Poisson or Bernoulli assumptions [47]. For the count endpoint (recurrent infections within 180 days) we complement these checks with the empirical frequency distribution, which highlights the heavy mass at zero and a long right tail characteristic of clustered clinical events. These views set expectations for subsequent modeling and ensure that any improvement in fit can be read directly against concrete patterns in the data.

Modeling proceeds in a deliberately incremental fashion. On the frequentist side, a Poisson regression serves as a baseline for the count outcome and is contrasted with a quasi–Poisson specification to estimate an overdispersion parameter while maintaining the mean structure [48]. To address extra zeros and heterogeneous variance jointly, a zero–inflated negative binomial (ZINB) model is then estimated, with the mean, dispersion, and zero–inflation components reported on epidemiologically familiar scales. In parallel, Bayesian formulations extend these ideas with priors over dispersion and inflation parameters, yielding posterior summaries that directly quantify uncertainty for coefficients and overdispersion: a zero–inflated beta–binomial (ZIBB) for grouped binary counts and a ZINB for recurrent events. Posterior computation uses gradient–based Markov chain Monte Carlo (HMC with the No–U–Turn Sampler), and adequacy is examined through convergence diagnostics and posterior predictive checks [46, 49, 50].

For transparency and comparability with prior chapters, all analyses are implemented in R and reported with threshold–free and coefficient–level summaries. Results are presented

in an aligned sequence: (i) coefficient comparisons for Poisson versus quasi–Poisson, (ii) conditional mean and inflation components for ZINB, and (iii) Bayesian posterior estimates for ZIBB and ZINB with dispersion parameters made explicit. These diagnostic figures motivate the shift from simple to enriched likelihoods, while the tabulated estimates show how allowing for overdispersion and zero–inflation stabilizes inference without sacrificing the clinical interpretability established in earlier GLM and hierarchical analyses [46–48].

2.7.1 Residuals in Generalized Linear Models

As an initial diagnostic for the independent–data logistic specification, we examined residual behavior on the probability scale to assess calibration and dispersion before introducing enriched likelihoods. The binned residual plot (Figure 2.32) summarizes mean residuals within quintiles of the fitted probabilities and compares them against ±2 standard–error bounds (for logistic regression model selected in Table 2.10). Residual means are largely centered around zero across the interior of the fitted range, with modest departures near the extremes, indicating mild lack of calibration where predicted risks are very low or very high; the absence of systematic trends through the center suggests no gross misspecification of the link or mean structure at this stage [47].

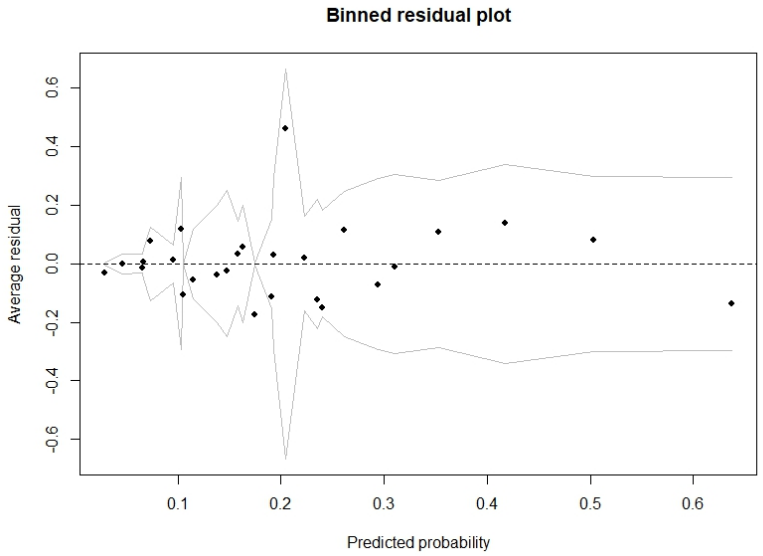


Figure 2.32: Binned residual plot from logistic regression. Each point represents the average residual within a bin of fitted values; the gray bounds correspond to ±2 SEs.

To complement this view with a simulation–based check less sensitive to the choice of residual scale, we inspected randomized quantile (DHARMa) residuals (Figure 2.33). The cloud is approximately symmetric with a few right–tail deviations and a slight increase in spread at higher fitted probabilities, a pattern compatible with residual overdispersion relative to the Bernoulli model.

These diagnostics motivate the transition in subsequent sections to models that accommodate extra–binomial variability and, for counts, to likelihoods that can also address zero–inflation, while retaining the same epidemiologic interpretation of coefficients [46, 47].

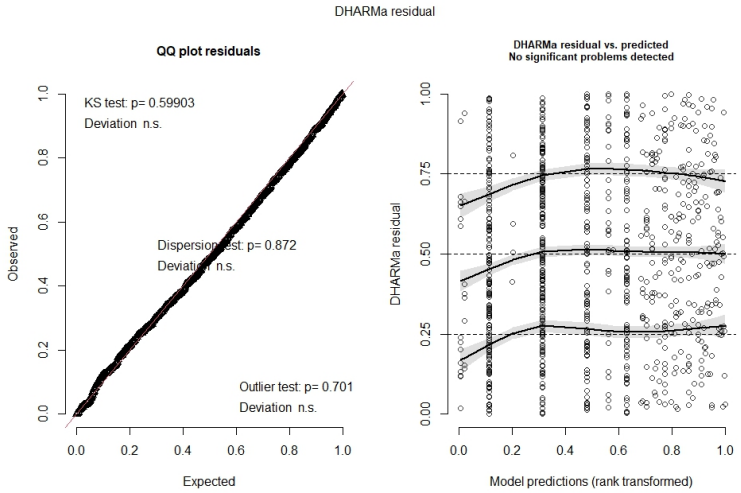


Figure 2.33: DHARMa residual plot from logistic regression.

2.7.2 Frequentist Models for Overdispersion

Binary Outcomes

Extending the residual diagnostics above to grouped binary data clarified when dispersion exceeds the binomial mean–variance relationship and standard errors require correction. We contrasted a single–trial logistic model (Over 1) with a proportional (grouped) binomial GLM (Over 2.0) in which each row aggregates multiple hospitalizations per subject. Whereas Over 1 showed no evidence of overdispersion, the grouped specification exhibited a dispersion statistic $\phi = 2.17$ ($p < 0.001$) and clear residual departures (Figure 2.34), consistent with unmodeled heterogeneity across trials within subjects. In epidemiologic applications where repeated hospital encounters differ in clinical context and risk, such extra–binomial variability is expected and, if ignored, can inflate test statistics and produce anticonservative p–values [47].

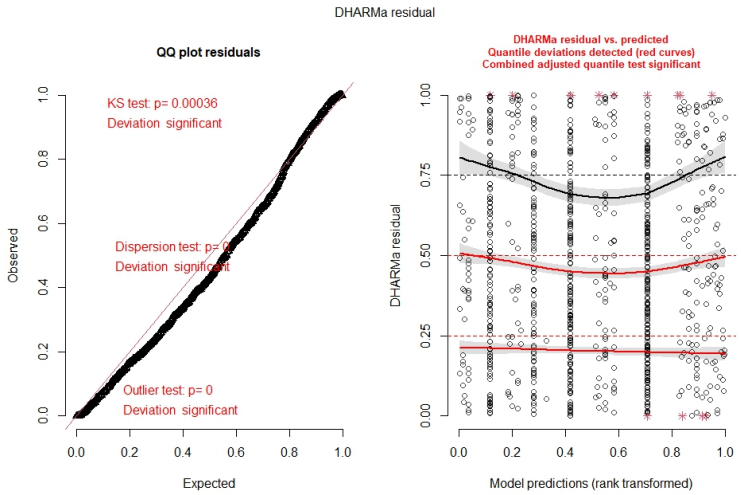


Figure 2.34: Residuals Check for Over 2.0 (grouped binomial GLM).

Table 2.35 reports the Over 2.0 coefficient estimates on the logit scale along with conventional (binomial) standard errors; several predictors appear significant under this naive variance. To adjust inference while preserving the mean structure, we refit the same model with a quasi–binomial specification (Over 2.1). As expected under quasi–likelihood, point estimates are identical, but standard errors inflate and some effects attenuate in significance (Table 2.36); for example, the catheter/drainage/dialysis term loses conventional significance once overdispersion is acknowledged. These results align with the general guidance that quasi–binomial models deliver calibrated uncertainty when grouped Bernoulli data depart from the fixed-variance assumption, without altering the fitted probabilities [48].

Table 2.35: Estimated coefficients from the model Over 2.0 (binomial)

Fixed Effect	**Estimate**	**Std. Error**	z **value**	**Pr**$(>\|z\|)$
Intercept	−1.778	0.073	−24.392	< 0.001
Sex at birth: Male	−0.364	0.091	−3.993	< 0.001
Race: White	−0.336	0.094	−3.561	< 0.001
Ethnicity: Other	−1.270	0.298	−4.257	< 0.001
Respiratory disorder	−0.324	0.322	−1.006	0.315
Endocrine disorder	−0.177	0.117	−1.512	0.131
Cancer	0.680	0.131	5.188	< 0.001
Catheter/Drainage/Dialysis	0.380	0.177	2.142	0.032
Intubation	1.517	0.273	5.552	< 0.001
Transplant	1.035	0.118	8.770	< 0.001

Table 2.36: Estimated coefficients from the model Over 2.1 (quasi-binomial)

Fixed Effect	**Estimate**	**Std. Error**	t **value**	**Pr**$(>\|t\|)$
Intercept	−1.778	0.108	−16.481	< 0.001
Sex at birth: Male	−0.364	0.135	−2.698	0.007
Race: White	−0.336	0.140	−2.406	0.016
Ethnicity: Other	−1.270	0.442	−2.876	0.004
Respiratory disorder	−0.324	0.477	−0.679	0.497
Endocrine disorder	−0.177	0.173	−1.022	0.307
Cancer	0.680	0.194	3.506	< 0.001
Catheter/Drainage/Dialysis	0.380	0.263	1.448	0.148
Intubation	1.517	0.404	3.751	< 0.001
Transplant	1.035	0.175	5.926	< 0.001

Count Outcomes

Extending the diagnostics to the count endpoint (number of infection–related encounters within 180 days), a Poisson GLM (Over 3.0) served as baseline. Residual checks showed a clear lack of fit with statistically significant dispersion—consistent with variance exceeding

the mean—and motivated variance–robust alternatives (Figure 2.35) [47]. The empirical frequency of events (Figure 2.36) displays a heavy mass at zero with a long right tail, indicating that extra–Poisson variation is accompanied by excess zeros in this cohort, a common pattern in surveillance data that warrants flexible mean–variance modeling [47].

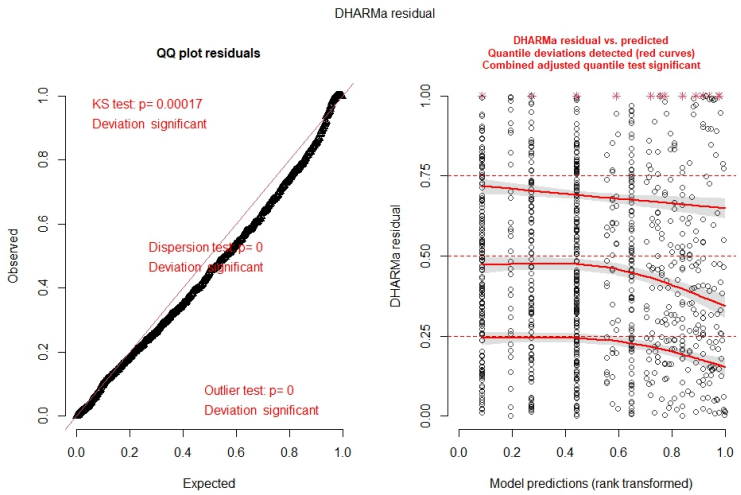


Figure 2.35: Residuals Check for Over 3.0.

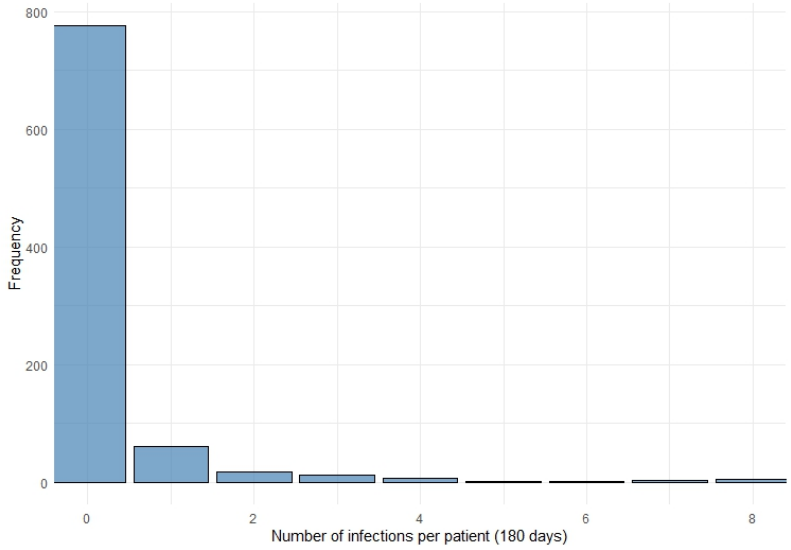


Figure 2.36: Frequency plot of recurrent infections per patient (180 days).

To calibrate uncertainty while preserving the Poisson mean structure, we refit the model using a quasi–Poisson specification (Over 3.1). As expected under quasi–likelihood, point estimates on the rate–ratio scale remain unchanged but confidence intervals widen (Table 2.37); the estimated dispersion parameter was $\hat{\phi} = 2.97$, confirming substantial overdispersion. This step provides a transparent correction for inflated test statistics without altering the substantive interpretation of coefficients [48].

Table 2.37: Comparison of estimated coefficients from Poisson and Quasi-Poisson models (Over 3.0 vs Over 3.1)

Fixed Effect	Poisson		Quasi-Poisson	
	Rate Ratio	95% CI	Rate Ratio	95% CI
Intercept	0.14	[0.11, 0.17]	0.14	[0.09, 0.23]
Sex at birth: Male	1.37	[1.05, 1.76]	1.37	[0.86, 2.20]
Race: White	0.68	[0.52, 0.89]	0.68	[0.43, 1.09]
Ethnicity: Other	1.04	[0.55, 1.95]	1.04	[0.34, 3.16]
Respiratory disorder	2.80	[1.88, 4.19]	2.80	[1.39, 5.63]
Endocrine disorder	2.88	[2.19, 3.79]	2.88	[1.80, 4.60]
Cancer	1.14	[0.79, 1.58]	1.14	[0.60, 2.15]
Cathet./Drainage/Dialysis	1.69	[1.00, 2.84]	1.69	[0.77, 3.97]
Intubation	3.52	[2.54, 4.89]	3.52	[2.03, 6.12]
Transplant	2.58	[1.50, 4.42]	2.58	[1.18, 5.59]

Given the prominent spike at zero, we next contrasted the quasi–Poisson with a zero–inflated negative binomial (ZINB) model (Over 3.2) that jointly addresses extra zeros and heterogeneous variance via a Poisson–Gamma mixture for the count component (negative binomial) and a logistic inflation component. For interpretability, the inflation sub-model included only an intercept, yielding an estimated log–odds of -1.46 and an implied structural–zero probability of $\hat{z} = \exp(-1.46) \approx 0.23$; conditional (count–model) rate ratios were broadly consistent with the quasi–Poisson, with modest differences reflecting improved accommodation of overdispersion and zeros (Table 2.38) [47].

Table 2.38: Conditional model estimates from Quasi-Poisson and ZINB models (Over 3.1 vs Over 3.2)

Fixed Effect	Quasi-Poisson		ZINB (count model)	
	Rate Ratio	95% CI	Rate Ratio	95% CI
Intercept	0.14	[0.09, 0.23]	0.16	[0.09, 0.30]
Sex at birth: Male	1.37	[0.86, 2.20]	1.39	[0.91, 2.11]
Race: White	0.68	[0.43, 1.09]	0.71	[0.47, 1.07]
Ethnicity: Other	1.04	[0.34, 3.16]	0.51	[0.13, 1.99]
Respiratory disorder	2.80	[1.39, 5.63]	2.94	[1.24, 6.99]
Endocrine disorder	2.88	[1.80, 4.60]	2.53	[1.63, 3.92]
Cancer	1.14	[0.60, 2.15]	1.42	[0.77, 2.61]
Cathet./Drainage/Dialysis	1.69	[0.77, 3.97]	1.95	[0.91, 4.17]
Intubation	3.52	[2.03, 6.12]	3.96	[2.23, 7.03]
Transplant	2.58	[1.18, 5.59]	3.01	[1.37, 6.59]

Residual diagnostics for ZINB indicate that dispersion and distributional shape are well controlled, though outlier tests remain positive and prompt targeted review (Figure 2.37).

These sequences from Poisson to quasi–Poisson to ZINB demonstrate how increasingly flexible likelihoods stabilize inference while retaining familiar epidemiologic effect scales.

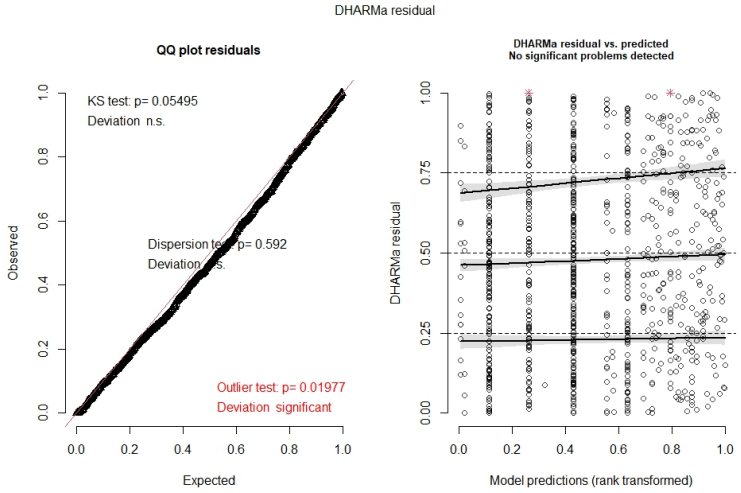


Figure 2.37: Residuals check for Over 3.2 (ZINB).

2.7.3 Bayesian Models for Overdispersion

Binary Outcomes

Motivated by the residual evidence of extra–binomial variation in grouped proportions, we fitted a zero–inflated beta–binomial (ZIBB) model (Over 4) to the proportion of infection–related diagnoses after hospitalization. The ZIBB formulation decomposes the outcome into a structural–zero process and a beta–binomial component that allows the success probability to vary across observations, thereby addressing both excess zeros and overdispersion within a single probabilistic framework [47]. Posterior computation proceeded via brms/Stan; convergence diagnostics were excellent with $\hat{R}$=1.00 for all parameters, indicating stable chains and reliable posterior summaries [46, 49].

Inference on the logit scale (Table 2.39) shows that endocrine disorder, cancer, catheter/drainage/dialysis, and intubation have 95% credible intervals excluding zero, while sex at birth, race, ethnicity, respiratory disorder, and transplant display posterior mass overlapping the null—an effect profile consistent with the multivariable GLM and hierarchical results reported earlier. The posterior for the dispersion (precision) parameter centers at ϕ=5.09 (95% CrI: 3.70–6.94), indicating moderate overdispersion relative to a fixed–probability binomial, and the zero–inflation component implies that approximately 8% of observations are structural zeros (95% CrI: 0%–20%), balancing heterogeneity with a modest excess of zeros in this cohort (Figure 2.38).

Count Outcomes

Guided by the excess zeros and extra–Poisson variation observed in the recurrent–infection counts, we estimated a Bayesian zero–inflated negative binomial (ZINB) model (Over 5) to quantify effects on the rate scale while propagating uncertainty in both dispersion and the

Table 2.39: Regression coefficients from the Bayesian ZIBB (Over 4) model

Fixed Effect	**Estimate**	**95% CrI**	$\hat{R}$
Intercept	−2.26	[−2.53, −1.96]	1.00
Sex at birth: Male	0.10	[−0.18, 0.38]	1.00
Race: White	−0.12	[−0.40, 0.17]	1.00
Ethnicity: Other	−0.11	[−0.96, 0.68]	1.00
Respiratory disorder	0.36	[−0.41, 1.14]	1.00
Endocrine disorder	0.71	[0.37, 1.04]	1.00
Cancer	0.62	[0.24, 1.01]	1.00
Cathet./Drainage/Dialysis	0.87	[0.29, 1.44]	1.00
Intubation	1.11	[0.42, 1.80]	1.00
Transplant	0.48	[−0.18, 1.07]	1.00

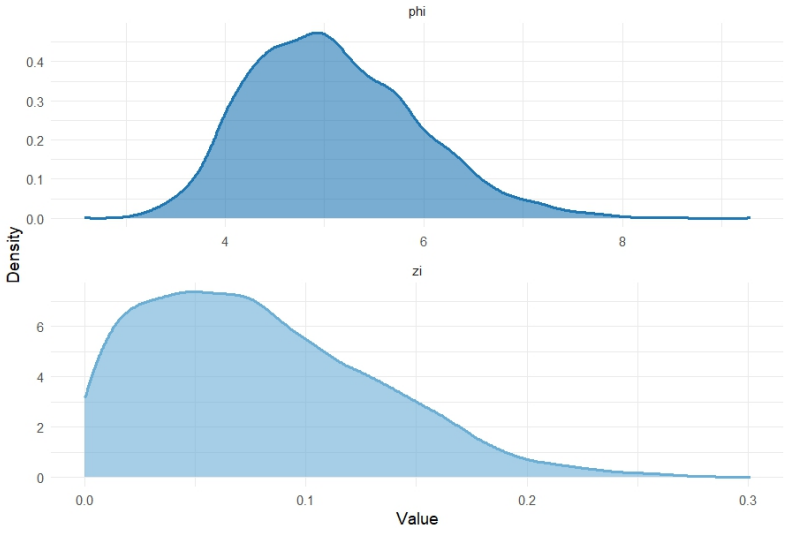


Figure 2.38: Above: Posterior distributions of the overdispersion (precision) parameter ϕ; below: Posterior distribution of the zero–inflation intercept for the Over 4 ZIBB model.

zero–inflation mechanism. The likelihood decomposes into (i) a logistic component for structural zeros and (ii) a negative–binomial count component with shape (overdispersion) parameter r, so that variance exceeds the mean in a data–driven manner; priors were weakly informative for coefficients and regularizing for r, and posterior simulation proceeded in brms/Stan with standard convergence checks [46, 49].

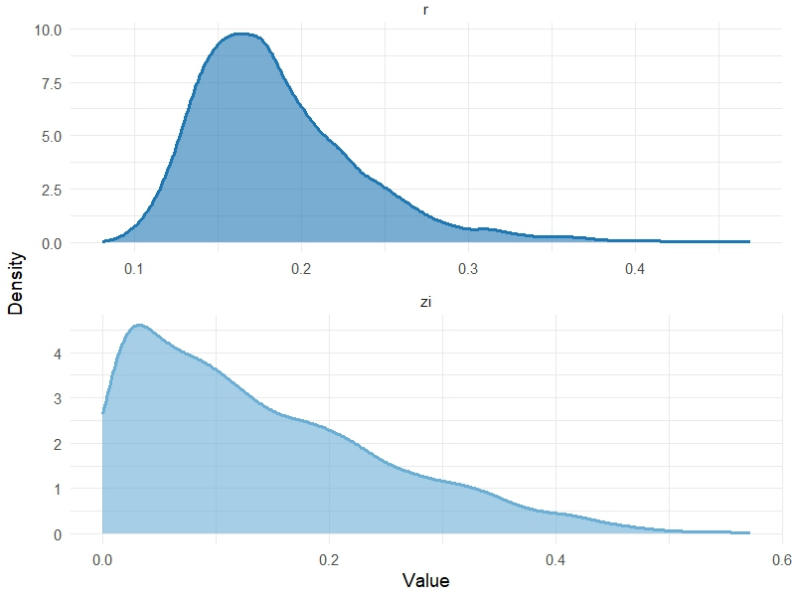


Figure 2.39: Above: Posterior Distributions of overdispersion parameter (shape r); below: Posterior Distributions of intercept from the Over 5 Bayesian ZINB model.

Figure 2.39 displays the posterior for r (top) and the zero–inflation intercept (bottom): the shape concentrates away from infinity (posterior mean $\hat{r} \approx 1.8$; 95% CrI $\approx [1.1, 2.7]$), confirming material overdispersion relative to Poisson, while the inflation intercept centers below zero (implying a structural–zero probability roughly one quarter of observations), consistent with the empirical spike at zero documented earlier. These parameter posteriors underpin coefficient inferences in Table 2.40, which we report on the incidence–rate–ratio (IRR) scale for epidemiologic interpretability [47].

The Bayesian ZINB results align with the quasi–Poisson and Frequentist ZINB comparisons but offer sharper, probability–based uncertainty summaries. Endocrine disorder (IRR 2.45, 95% CrI 1.67–3.73) and respiratory disorder (IRR 2.98, 1.22–7.12) show strong positive associations with recurrent events; procedure–related exposures remain prominent, with intubation (IRR 3.92, 2.24–7.18) and transplant (IRR 2.94, 1.41–6.66) elevated, and catheter/drainage/dialysis trending upward but imprecise. Demographic terms are near null once clinical covariates are included.

2.8 Generalized Linear Models with Missing Data

Building on the independent, correlated, and zero–inflated analyses developed in earlier sections, we now address missing data as a first–order property of real–world clinical

Table 2.40: Regression coefficients from the Bayesian ZINB (Over 5) model

Predictor (count model)	**IRR (median)**	**95% CrI**	$\hat{R}$
Intercept	0.16	[0.09, 0.30]	1.00
Sex at birth: Male	1.39	[0.95, 2.04]	1.00
Race: White	0.71	[0.49, 1.04]	1.00
Ethnicity: Other	0.52	[0.12, 1.92]	1.00
Respiratory disorder	2.98	[1.22, 7.12]	1.00
Endocrine disorder	2.45	[1.67, 3.73]	1.00
Cancer	1.41	[0.77, 2.55]	1.00
Cathet./Drainage/Dialysis	1.93	[0.93, 4.08]	1.00
Intubation	3.92	[2.24, 7.18]	1.00
Transplant	2.94	[1.41, 6.66]	1.00
Zero–inflation (intercept)	logit scale ≈ -1.47	(P(structural zero) ≈ 0.23)	1.00
Overdispersion (shape r)	posterior mean ≈ 1.8	(95% CrI $\approx$[1.1, 2.7])	1.00

cohorts. Incomplete covariates and outcomes arise from workflow, access, and documentation patterns; treating the observed records as if they were complete can bias effect estimates, understate uncertainty, and distort model comparison. A principled analysis therefore begins by articulating the missingness mechanism—*missing completely at random* (MCAR), *missing at random* (MAR), or *missing not at random* (MNAR)—and by aligning estimation with assumptions that are credible for the data–collection process [24]. Under MCAR, complete–case analyses are unbiased but inefficient; under MAR, likelihood–based or multiple–imputation strategies are required for valid inference; under MNAR, the joint data–generating process must be modeled explicitly [25].

Multiple Imputation by Chained Equations (MICE) supplies a practical route under MAR by repeatedly imputing the incomplete fields with conditionally specified models, fitting the analysis model to each completed dataset, and combining estimates with Rubin's rules; this approach recovers information lost to missingness while transparently propagating imputation uncertainty [26, 27]. In epidemiologic workflows, flexible conditional models (e.g., logistic, ordinal, and count regressions) allow MICE to respect data types and constraints, and auxiliary variables can improve the plausibility of MAR by capturing predictors of missingness [69]. The R implementation (`mice`) operationalizes these ideas with diagnostics for convergence and imputation quality, facilitating reproducible pipelines from imputation through model estimation and pooling [70].

A Bayesian perspective extends these principles by treating missing values as unknown parameters and drawing jointly from the posterior of model coefficients and imputations; this unifies imputation and analysis, yields full posterior predictive distributions, and can accommodate MNAR through explicit modeling of the missingness process [28]. In practice, Bayesian MICE (imputation with congenial priors and pooling) and fully Bayesian joint models (FBJM) both propagate uncertainty, but the latter integrates all components in a single generative specification, which is especially useful when the outcome model is complex or the missingness mechanism is structurally informative [26].

To maintain continuity with prior sections, we analyze the same All of Us cohort and

binary endpoint (infection within 180 days), report effects on epidemiologically familiar scales, and pair model summaries with visual explanations of MCAR, MAR, and MNAR to ground assumptions in study design rather than post hoc diagnostics alone. These imputation–based and fully Bayesian analyses provide a coherent framework in which uncertainty from missingness is quantified explicitly, and substantive conclusions are reconciled across mechanisms and modeling choices [27, 69].

2.8.1 Underlying Causes of Missing Data

Interpretable handling of incomplete records begins by specifying a credible mechanism for how values go missing in our cohort. The three canonical mechanisms MCAR, MAR, and MNAR differ in the role played by observed and unobserved quantities and, in turn, determine which estimators yield unbiased inference and how uncertainty must be propagated [24]. In practice, mechanism choice is guided by study operations, data–capture patterns, and auxiliary variables, with graphical summaries used to anchor assumptions in data features rather than post hoc convenience [27]. To evaluate model behavior under controlled missingness, we artificially induced MCAR, MAR, and MNAR patterns.

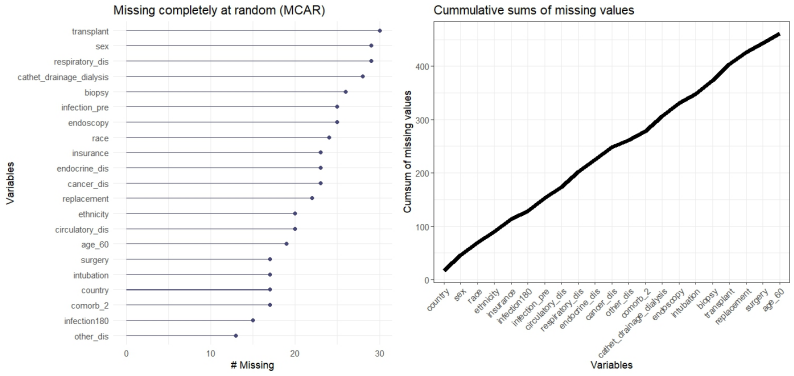


Figure 2.40: Missing Completely at Random (MCAR).

Under MCAR, the probability that a value is missing does not depend on either observed or unobserved data. Complete–case analyses remain unbiased, albeit inefficient, and simple diagnostics—such as comparing observed covariate distributions between missing and non–missing groups—should not reveal systematic differences beyond sampling fluctuation [25]. In our setting, MCAR would correspond to sporadic system outages or truly random omissions in intake forms. Figure 2.40 provides a schematic that separates the missingness process from both the covariates and the outcome, emphasizing that the mechanism is ignorable for likelihood–based inference but still costs information [26].

MAR allows missingness to depend on *observed* variables but not on the missing values themselves, conditional on what has been recorded. This mechanism matches many clinical workflows: for example, lab data may be absent more often in younger, low–acuity patients, even if the unobserved lab values would not differ further once age and acuity are conditioned upon. Valid inference then requires modeling that conditions on those predictors of missingness—either directly via likelihood or through multiple imputation by chained equations (MICE), which replaces missing entries with draws from their conditional distributions and combines estimates across completed datasets [27]. Figure 2.41

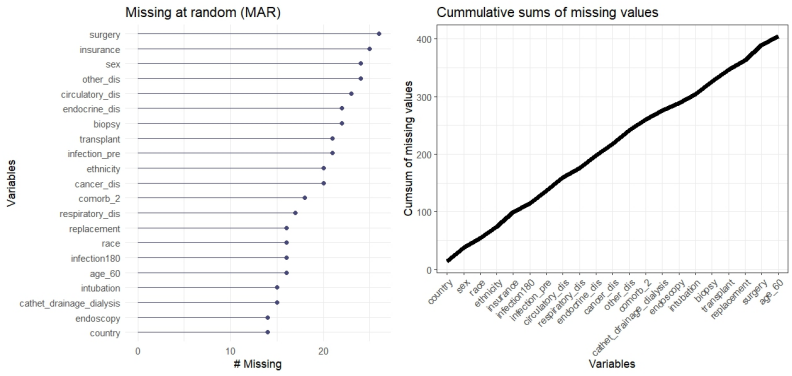


Figure 2.41: Missing At Random (MAR).

highlights the role of observed information in rendering the missingness ignorable; the analysis model and imputation models should be *congenial* to avoid incompatibilities that bias pooled estimates [69].

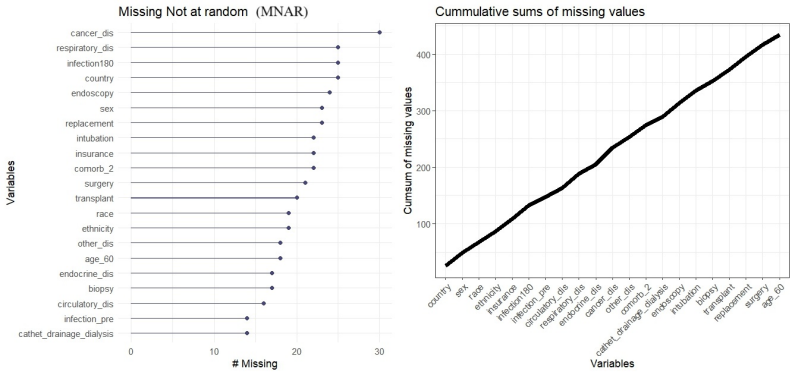


Figure 2.42: Missing Not at Random (MNAR).

MNAR arises when the probability of missingness depends on the unobserved value even after conditioning on observed data (e.g., follow–up cultures are more likely to be absent precisely when infection burden is unusually high or unusually low). In this case, neither complete–case analysis nor MAR–based imputation is guaranteed to be unbiased. Identification requires structure: sensitivity–analysis models (selection, pattern–mixture, or shared–parameter) or fully Bayesian joint models that specify the data–generating process for outcomes, covariates, and missingness together, with priors encoding clinically plausible departures from MAR [26]. Figure 2.42 depicts the direct path from the unobserved value to the missingness indicator, clarifying why assumptions must be made explicit and interrogated in sensitivity analyses [27].

2.8.2 Filling in the Blanks: Practical Tools for Missing Values

Bridging the mechanism taxonomy to analysis, we next contrasted three pragmatic routes for incomplete records in the same All of Us cohort. In the complete–case fit, only rows without any missing entries in analysis variables are retained; the approach is unbiased only under MCAR and typically sacrifices power through reduced n [26]. Under

MAR, we imputed with chained equations and combined estimates via Rubin's rules to recover efficiency and propagate uncertainty; congenial imputation models and auxiliary predictors safeguard validity [27, 70]. When MNAR is plausible, identification requires structure—either explicit sensitivity models (selection/pattern–mixture) or fully Bayesian joint models that treat the missing values as unknowns and draw them jointly with parameters, yielding posterior summaries that directly encode uncertainty in the missingness process [26, 28]. In what follows, we report the complete–case GLM as the natural point of departure and then motivate movement toward MAR and MNAR frameworks where patterns in the data and study operations indicate departures from MCAR.

List–wise Deletion (Complete-case GLM).

Applied to the working dataset, complete–case analysis delivered stable coefficients for several clinical predictors but, as expected, depends critically on the MCAR assumption; when missingness correlates with recorded or unrecorded characteristics, estimates can be biased and standard errors optimistic because informative patterns are discarded [24]. In our cohort, the complete–case GLM under three simulated mechanisms (MCAR, MAR, MNAR) shows how effect magnitudes and p–values shift once missingness ceases to be purely random (Table 2.41). Notably, the MCAR column reflects relatively precise associations for respiratory and endocrine disease, whereas MAR and MNAR columns display attenuation or instability for demographic terms, consistent with nonrandom loss of information under those mechanisms. These patterns reinforce that list–wise deletion is tenable only when MCAR is credible, and otherwise serves merely as a benchmark for more principled strategies.

Table 2.41: Comparison of GLM Results under MCAR, MAR, and MNAR Missingness

	Odds Ratio			**95% CI**			***p*-value**		
Variable	**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**
(Intercept)	0.31	1.28	2.50	[0.05, 1.93]	[0.02, 0.85]	[0.03, 1.73]	0.219	0.035	0.166
Country of Birth (ref: Other)									
USA	0.81	0.37	1.03	[0.33, 2.04]	[0.14, 0.98]	[0.40, 8.22]	0.640	0.043	0.956
Sex (ref: Female)									
Male	1.21	1.49	1.17	[0.63, 2.34]	[0.69, 3.23]	[0.61, 2.22]	0.559	0.332	0.632
Other	0.67	8.73	8.64	[0.55, 3.11]	[0.04, 7.17]	[0.04, 7.17]	0.734	0.872	0.906
Age (ref: ≤60)									
>60	1.37	1.04	1.58	[0.65, 2.84]	[0.44, 2.32]	[0.77, 3.21]	0.304	0.933	0.207
Race (ref: Black or African Americans)									
Other	0.61	0.38	0.43	[0.21, 1.73]	[0.24, 0.99]	[0.15, 1.21]	0.356	0.048	0.105
White	0.49	9.23	6.09	[0.21, 1.18]	[0.38, 2.31]	[0.27, 1.43]	0.110	0.861	0.240
Ethnicity (ref: Hispanic or Latino)									
Others	0.59	1.83	4.24	[0.20, 1.65]	[0.53, 6.08]	[0.13, 1.29]	0.317	0.327	0.136
Health Insurance (ref: No)									
Yes	0.61	2.20	1.83	[0.20, 2.25]	[0.25, 19.42]	[0.33, 10.03]	0.420	0.668	0.487
Previous Infection History (ref: No)									
Yes	1.27	6.75	0.94	[0.58, 2.69]	[0.22, 18.06]	[0.42, 2.01]	0.536	0.462	0.885

2.8.3 Multiple Imputation and the MICE Algorithm

To translate the mechanism assumptions into analysis, MICE was implemented under MCAR, MAR, and MNAR scenarios using fully conditional specification with iterated conditional models for each incomplete variable, followed by Rubin's rules to pool estimates and standard errors across m completed datasets [26, 27, 69, 70]. In our workflow, imputation and pooling proceeded as in R's `mice` pipeline, with model fitting repeated on each imputed dataset and the between– and within–imputation variability combined to yield final inference; conceptually, the MICE algorithm cycles through variables, draws imputations from their conditional models, and repeats until stabilization before generating multiple completed datasets for analysis, ensuring that uncertainty from missingness is propagated to coefficient estimates and their variances [24]. The analytic target remained a multivariable GLM for infection within 180 days, preserving epidemiologic interpretability while addressing information loss due to missingness. :

Pooled results in Table 2.42 show the expected pattern: relative to complete–case fits, MICE yields more stable estimates and standard errors, with modest attenuation of several covariate effects under MAR and MNAR where missingness relates to observed or unobserved predictors of inclusion. Clinical signals such as circulatory disease, respiratory disease, endocrine disease, intubation, transplant, and replacement remain directionally consistent and largely significant across mechanisms, whereas demographic terms are near null after adjustment and exhibit sensitivity to the assumed missingness mechanism. This behavior is coherent with the mechanism taxonomy: when missingness is ignorable after conditioning on observed data (MAR), pooling improves precision and calibrates uncertainty; when missingness depends on unobserved values (MNAR), standard errors widen and some effects weaken, reflecting additional uncertainty that MICE can only partially resolve without explicit MNAR structure. These comparisons, read alongside the mechanism definitions, support the use of MICE as a pragmatic default under MAR with transparent reporting of pooled estimates and their variability.

2.8.4 Bayesian Multiple Imputation by Chained Equations

As a Bayesian extension of chained equations, Bayesian MICE augments each conditional imputation model by sampling its parameters from the posterior at every iteration and then drawing missing values from the corresponding predictive distribution; this preserves the familiar FCS/MICE workflow while propagating uncertainty from both parameters and imputations [26]. In practice, the algorithm cycles through variables with missingness, iteratively updating $(\theta_j, Y_j^{\text{mis}})$ via posterior draws, and repeats the full cycle m times to generate multiple completed datasets for analysis and pooling [27]. In settings with moderate sample sizes or collinearity, the use of priors regularizes conditional models and stabilizes convergence relative to classical MICE, while leaving the epidemiologic interpretation of downstream GLMs intact.

Applied to the infection endpoint, posterior summaries from Bayesian logistic regressions fitted to the imputed datasets under MCAR, MAR, and MNAR show a coherent pattern across mechanisms (Table 2.43). The coefficient for endocrine disease is consistently positive and precise (posterior mean $\approx$ 1.11–1.12 with 95% CrIs excluding 0), and the age $>$ 60 effect remains positive with modest variability ($\approx$ 0.26–0.30), while intercept shifts are small across mechanisms; all chains exhibit excellent convergence ($\hat{R} = 1.00$)

Table 2.42: Pooled GLM Estimates after MICE Imputation under MCAR, MAR, and MNAR

Variable	Estimate			Std. Error			P-value		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR
(Intercept)	-1.81	-1.55	-1.81	0.68	0.70	0.67	0.008	0.027	0.007
Country of Birth (ref: Other)									
USA	-0.23	-0.50	-0.13	0.35	0.36	0.37	0.506	0.170	0.723
Sex (ref: Female)									
Male	0.33	0.38	0.40	0.23	0.23	0.24	0.156	0.193	0.327
Other	-0.78	-0.74	-0.25	1.12	1.12	1.08	0.478	0.508	0.815
Age (ref: $\leq$60)									
>60	1.06	0.12	0.05	0.47	0.26	0.27	0.036	0.637	0.828
Race (ref: Black or African American)									
Other	-0.26	-0.25	-0.31	0.40	0.42	0.40	0.520	0.548	0.442
White	-0.25	-0.13	-0.21	0.31	0.32	0.33	0.424	0.673	0.529
Ethnicity (ref: Hispanic or Latino)									
Other	-0.33	-0.34	-0.51	0.39	0.41	0.39	0.384	0.407	0.198
Health Insurance (ref: No)									
Yes	-0.27	-0.25	-0.20	0.45	0.44	0.49	0.555	0.552	0.632
Previous Infection History (ref: No)									
Yes	-0.01	-0.24	0.03	0.28	0.29	0.29	0.981	0.407	0.944
Circulatory Disease (ref: No)									
Yes	1.28	1.40	2.09	0.50	0.57	0.56	0.010	0.017	0.032
Respiratory Disease (ref: No)									
Yes	1.78	1.39	1.78	0.53	0.57	0.52	<0.001	0.031	0.005
Endocrine Disease (ref: No)									
Yes	1.31	1.29	1.29	0.30	0.30	0.30	<0.001	<0.001	0.025
Cancer (ref: No)									
Yes	1.06	1.03	1.06	0.39	0.38	0.38	0.009	0.007	0.007
Other Disease (ref: No)									
Yes	1.32	1.30	1.31	0.77	0.76	0.81	0.094	0.092	0.098
Multiple Comorbidity (ref: No)									
Yes	-1.63	-1.51	-1.55	0.56	0.57	0.58	0.004	0.009	0.008
Catheter Drainage (ref: No)									
Yes	0.57	0.34	0.33	0.50	0.55	0.55	0.259	0.562	0.588
Endoscopy (ref: No)									
Yes	-0.40	-0.38	-0.43	0.62	0.58	0.59	0.508	0.530	0.509
Intubation (ref: No)									
Yes	1.38	1.41	1.39	0.39	0.40	0.38	0.001	<0.001	<0.001
Biopsy (ref: No)									
Yes	0.84	0.82	0.80	0.48	0.49	0.48	0.088	0.095	0.099
Transplant (ref: No)									
Yes	1.22	1.25	1.25	0.60	0.58	0.58	0.044	0.037	0.036
Replacement (ref: No)									
Yes	-1.12	-1.21	-1.27	0.56	0.55	0.54	0.036	0.028	0.021
Surgery (ref: No)									
Yes	0.61	0.72	-0.01	0.47	0.46	0.38	0.188	0.126	0.982

Table 2.43: Posterior Summaries from Bayesian Logistic Regression after MICE Imputation

Dataset	Parameter	Estimate	Est. Error	2.5% CI	97.5% CI	Rhat	ESS (Bulk / Tail)
MCAR	Intercept	-2.33	0.06	-2.45	-2.22	1.00	3364 / 2459
	Age (ref: ≤60) > 60	0.26	0.07	0.12	0.40	1.00	3841 / 2668
	Endocrine Disease (ref: No) Yes	1.12	0.09	0.93	1.29	1.00	3496 / 2702
MAR	Intercept	-2.35	0.06	-2.46	-2.24	1.00	2939 / 2514
	Age (ref: ≤60) > 60	0.30	0.07	0.16	0.44	1.00	3234 / 2723
	Endocrine Disease (ref: No) Yes	1.12	0.09	0.95	1.29	1.00	3201 / 2790
MNAR	Intercept	-2.37	0.06	-2.48	-2.26	1.00	3193 / 2610
	Age (ref: ≤60) > 60	0.29	0.07	0.14	0.43	1.00	3375 / 2674
	Endocrine Disease (ref: No) Yes	1.11	0.09	0.93	1.28	1.00	3213 / 2798

and large bulk/tail effective sample sizes. Relative to the classical MICE results, credible intervals are slightly narrower and estimates more stable, particularly under MAR/MNAR—an expected benefit of posterior regularization and full uncertainty propagation through the imputation step. Read alongside the pooled GLM estimates from the previous subsection, these results indicate that Bayesian MICE preserves the substantive signal while sharpening uncertainty quantification in a principled way.

2.8.5 Fully Bayesian Joint Modeling (FBJM)

To address settings where the probability that a value is missing may depend on the unobserved value itself, we estimated fully Bayesian joint models that unite the outcome model, the covariate model, and the missingness mechanism in a single generative specification. In this framework, missing entries are treated as unknown parameters and are sampled alongside regression coefficients and hyperparameters, so that uncertainty about missingness is propagated coherently into the target inferences. For the binary outcome (infection within 180 days), the analysis model was a logistic GLM with the same clinically motivated predictors used throughout; the missingness mechanism was represented with selection–model terms, and weakly informative priors were placed on regression coefficients with regularizing priors on variance components [46, 49]. Conceptually, FBJM complements (Bayesian) MICE by replacing chained conditional models with a single joint model, which is particularly valuable when MNAR cannot be ruled out and identification rests on transparent, clinically defensible assumptions [26, 28].

Posterior summaries under MCAR, MAR, and MNAR are presented in Table 2.44. The signal for endocrine disease remains robustly positive across mechanisms with credible intervals excluding zero, in agreement with earlier GLM, hierarchical, and imputation results. In contrast, the age >60 coefficient is positive but its 95% credible intervals straddle zero under MAR and MNAR, reflecting the additional uncertainty introduced when the missingness process is modeled explicitly rather than assumed ignorable. Intercepts are stable across mechanisms, and all chains exhibit excellent convergence ($\hat{R} \approx 1.00$) with large bulk and tail effective sample sizes. In summary, the Bayesian–MICE estimates,

Table 2.44: Posterior Summaries from Fully Bayesian Joint Modeling (FBJM) under MCAR, MAR, and MNAR. *Tail Prob.* denotes the posterior probability that the parameter is above 0 (for a positive effect) or below 0 (for a negative effect). *GR-crit* refers to the Gelman–Rubin convergence statistic $\hat{R}$, which being near 1.00 indicates convergence.

Dataset	Parameter	Mean	SD	2.5%	97.5%	Tail Prob.	GR-crit	MCE/SD
MCAR	Intercept	-2.2931	0.148	-2.591	-2.010	0.000	1.00	0.0145
	Age (ref: ≤60) > 60	0.0914	0.228	-0.362	0.540	0.684	1.00	0.0146
	Endocrine Disease (ref: No) Yes	1.1693	0.233	0.711	1.630	0.000	1.00	0.0139
MAR	Intercept	-2.2924	0.143	-2.579	-2.016	0.000	1.00	0.0142
	Age (ref: ≤60) > 60	0.0902	0.229	-0.362	0.532	0.684	1.00	0.0133
	Endocrine Disease (ref: No) Yes	1.1698	0.231	0.706	1.615	0.000	1.00	0.0127
MNAR	Intercept	-2.3640	0.147	-2.664	-2.086	0.000	1.00	0.0143
	Age (ref: ≤60) > 60	0.1380	0.232	-0.317	0.594	0.544	1.00	0.0128
	Endocrine Disease (ref: No) Yes	1.1850	0.236	0.723	1.655	0.000	1.00	0.0137

FBJM yields broadly consistent conclusions for well–identified predictors and provides principled, posterior–based widening of intervals where sensitivity to missingness assumptions is material [26, 27].

2.9 Interrupted Time Series Model in Infectious Disease Research and Surveillance

Extending the cohort–based regression and hierarchical analyses to population–level dynamics, we evaluated the impact of a national stewardship campaign on routine prescribing using interrupted time series (ITS). The outcome is the monthly proportion of antibiotic prescriptions among patients with viral diagnoses in the All of Us data from January 2010 through December 2022 (156 observations), with the *Be Antibiotics Aware* (BAA) campaign in November 2017 marked as the intervention point. Time series of prescribing display secular trends, strong annual seasonality, and irregular fluctuations typical of respiratory illness patterns; these structures motivate segmented regression to estimate immediate level changes and post–intervention slope changes, and ARIMA–class models to account for autocorrelation and seasonality so that intervention effects are not confounded by serial dependence [15–17]. The ITS framework is read on epidemiologically interpretable scales—changes in level and trend of inappropriate prescribing—while diagnostics assess linearity, homoscedasticity, and residual autocorrelation to support credible counterfactual contrasts [18–20].

Inference proceeds in parallel frequentist and Bayesian strands to enhance robustness and interpretability. A classical segmented OLS model provides transparent estimates of pre–campaign trend, immediate post–campaign level shift, and change in slope, with residual checks guiding the addition of autoregressive or moving–average components. Seasonal ARIMA specifications then capture the dominant serial structure so that intervention coefficients reflect changes beyond predictable seasonality; model selection is

based on information criteria and residual whiteness [15]. Bayesian counterparts extend these ideas by placing priors on ARIMA parameters and regression effects and by summarizing uncertainty via posterior distributions, with computation based on HMC/NUTS and convergence assessed through rank–normalized $\hat{R}$ and effective sample sizes [46, 49, 50]. Finally, a Bayesian structural time series (BSTS) decomposes the series into trend, seasonal, and regression (intervention) components within a state–space formulation, yielding posterior inclusion and contribution estimates that link qualitative component behavior with quantitative effect sizes on the prescribing scale. Forecast comparisons—without and with the intervention encoded—anchor the interpretation of estimated effects in observable counterfactual trajectories, and uncertainty bands reflect both parameter and disturbance variance rather than sampling noise alone [15, 18].

Together with the segmented regression, seasonal ARIMA, Bayesian ARIMA, and BSTS models provide convergent evidence on the magnitude and persistence of the BAA campaign's association with prescribing practice. The analysis is organized so that each approach answers the same substantive question through slightly different assumptions about serial dependence and prior information, with figures reporting observed series, fitted trends, residual diagnostics, and out–of–sample forecasts, and tables presenting level and slope changes with corresponding uncertainty on scales meaningful for surveillance and stewardship policy [17, 19, 46].

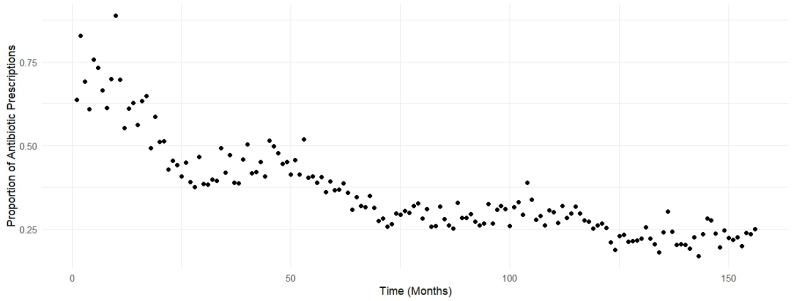


Figure 2.43: Monthly proportion of antibiotic prescriptions among patients with viral diagnoses.

Figure 2.43 displays the monthly time series of the proportion of antibiotic prescriptions among patients with viral diagnoses over the full study period. This raw plot highlights several key features that are characteristic of time series data and warrant closer examination. First, there is a clear long-term downward trend, suggesting that antibiotic prescribing has been decreasing steadily over time. Second, the series exhibits irregular fluctuations, with short-term increases and decreases that do not follow a strictly linear pattern. These may reflect external influences, seasonal variation, or noise. Third, the variability in the early part of the series appears greater than in later periods, indicating potential changes in variance or structural breaks over time.

2.9.1 Application of Classical ARIMA Models in ITS Analysis

To complement segmented regression and isolate the intervention signal from serial dependence, we fitted seasonal ARIMA models to the monthly prescribing series and selected

the specification by minimizing AIC_c over a grid with Fourier seasonal terms. The selected model, $\mathrm{ARIMA}(0,1,1)(1,0,0)_{12}$ with drift and two harmonic regressors ($S_{1\text{–}12}$, $C_{1\text{–}12}$), implies first–difference stationarity, a short–memory MA(1) component, and a seasonal AR(1) at period 12; Table 2.45 summarizes the fit. Model diagnostics and fit indices indicate an adequate representation of the data: log–likelihood = 119.42, AIC = −226.83, AIC_c = −226.26, BIC = −208.57, residual variance $\widehat{\sigma}^2$ = 0.01284, RMSE = 0.0457, MAE = 0.0330, and residual ACF at lag 1 of 0.0143, all consistent with well–behaved residuals for ITS inference.

Table 2.45: Model fit statistics and training-set error measures

Metric	Value
σ^2	0.01
Log-likelihood	119.42
AIC	−226.83
AICc	−226.26
BIC	−208.57
ME	−0.01
RMSE	0.05
MAE	0.03
MPE	−1.22
MAPE	8.98
MASE	0.51
ACF1	0.01

Coefficient estimates and 95% confidence intervals are reported in Table 2.46. The nonseasonal MA(1) term is strongly negative (estimate −0.6376, $p < 0.001$), capturing short–term error correction, while the seasonal AR(1) is modestly negative (estimate −0.2655, $p = 0.0019$), consistent with annual persistence in departures from trend. The drift is small but significantly negative (estimate −0.0073, $p = 0.0057$), reflecting a gradual downward tendency in the differenced series. By contrast, the included Fourier harmonics are not statistically distinguishable from zero once the seasonal AR term is present, indicating that the core SARIMA structure subsumes most seasonal variation.

Table 2.46: Coefficient estimates and 95% confidence intervals for the $\mathrm{ARIMA}(0,1,1)(1,0,0)_{12}$ model

Term	**Estimate**	**Std. Error**	***z*-value**	***p*-value**	**2.5% CI**	**97.5% CI**
MA1	−0.6376	0.0621	−10.26	< 0.001	−0.7594	−0.5158
SAR1	−0.2655	0.0854	−3.10	0.0019	−0.4331	−0.0976
Drift	−0.0073	0.0026	−2.77	0.0057	−0.0125	−0.0021
S1–12	0.0027	0.0109	0.25	0.8024	−0.0187	0.0241
C1–12	−0.0115	0.0108	−1.07	0.2857	−0.0327	0.0096

Residual checks (Figure 2.44) show no visible trend or seasonality in the residual series, and residual autocorrelations lie within 95% limits; the Ljung–Box statistic

($Q^* = 22.82$, df = 22, $p = 0.4118$) fails to reject white noise, and the histogram aligns closely with normality, supporting the validity of standard errors and forecast bands under the Gaussian state–space approximation [15]. These diagnostics, together with the information–criterion profile, justify carrying the ARIMA$(0,1,1)(1,0,0)_{12}$ specification forward for counterfactual forecasting and for comparison with Bayesian ARIMA in the next subsection. :contentReference[oaicite:2]index=2

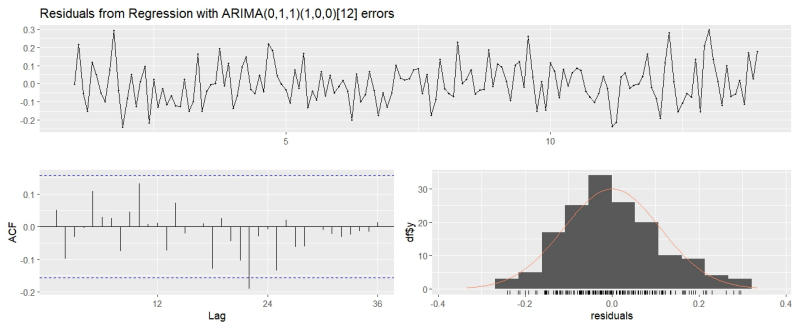


Figure 2.44: Residual diagnostics for the ARIMA$(0,1,1)(1,0,0)_{12}$ model: residual time plot (top), ACF (bottom-left), and histogram with normal density (bottom-right).

2.9.2 Application of Bayesian ARIMA Models in ITS Analysis

To complement the classical SARIMA fit and express serial dependence with full probabilistic uncertainty, we estimated a Bayesian ARIMA model with the same differencing and seasonal structure as the selected frequentist specification, namely ARIMA$(0,1,1)(1,0,0)_{12}$ with drift. Weakly informative priors were placed on the moving–average coefficient, the seasonal autoregression, the drift (local linear trend in the differenced series), and the innovation scale. Posterior inference used Hamiltonian Monte Carlo with the No–U–Turn Sampler, and convergence was assessed via rank–normalized $\hat{R}$ and effective sample sizes; all chains mixed well with $\hat{R} \approx 1.00$ and large bulk/tail ESS, indicating stable estimation [46, 49, 50]. Summary results (Table 2.47) show posterior means close to the classical estimates, with a strongly negative MA(1) (posterior mean ≈ -0.61), a modest negative seasonal AR(1) at period 12 (posterior mean ≈ -0.18), and a small negative drift (posterior mean ≈ -0.0034), all consistent with a gradually declining differenced process whose short–memory and annual components are well captured [15].

Figure 2.45 displays trace and density plots for key parameters (μ_0, σ_0, and ma[1]). The traces exhibit rapid mixing without stickiness or drift, and the posterior densities are unimodal and well–separated from prior support boundaries, supporting reliable uncertainty quantification for subsequent forecasting and intervention contrasts. Residual diagnostics in Figure 2.46 mirror the classical checks: residuals appear mean–zero with no visible seasonal pattern; autocorrelations fall within simulation bands, and the histogram aligns closely with a Gaussian approximation. Read alongside the frequentist SARIMA, the Bayesian fit thus provides a coherent, distributional summary of serial structure—tight posteriors for the MA and seasonal AR terms and a small, negative drift—while keeping the ITS interpretation intact for level and slope changes around the stewardship campaign [15, 46].

Table 2.47: Bayesian SARIMA(0,1,1)(1,0,0)[12].reg[8]: Posterior summary (mean, SE, central 90% interval), effective sample size (ESS), and $\hat{R}$.

Parameter	Mean	SE	5%	95%	ESS	$\hat{R}$
mu0	−0.0034	0.0000	−0.0059	−0.0008	5789.544	1.0002
sigma0	0.0481	0.0000	0.0436	0.0531	5909.916	1.0005
ma	−0.6097	0.0008	−0.7032	−0.4982	5735.574	0.9999
sar	−0.1768	0.0011	−0.3137	−0.0418	6051.127	1.0003
breg.1	−0.0008	0.0001	−0.0110	0.0093	6256.557	1.0002
breg.2	−0.0050	0.0001	−0.0147	0.0048	5877.158	1.0001
breg.3	−0.0069	0.0001	−0.0149	0.0011	5931.017	1.0001
breg.4	−0.0024	0.0001	−0.0104	0.0054	5395.378	1.0000
breg.5	0.0040	0.0001	−0.0034	0.0115	6013.274	1.0003
breg.6	−0.0060	0.0001	−0.0143	0.0007	6024.429	1.0009
breg.7	−0.0050	0.0001	−0.0125	0.0025	5890.100	0.9999
breg.8	−0.0059	0.0001	−0.0132	0.0015	5809.507	0.9999
loglik	251.1472	0.0341	246.2714	254.7422	5997.996	1.0001

Notes: Model: $y \sim \text{SARIMA}(0,1,1)(1,0,0)_{[12]}$ with 8 regression terms. 156 observations (current: 155). Samples drawn via NUTS; ESS denotes effective sample size; $\hat{R}$ is the potential scale-reduction factor.

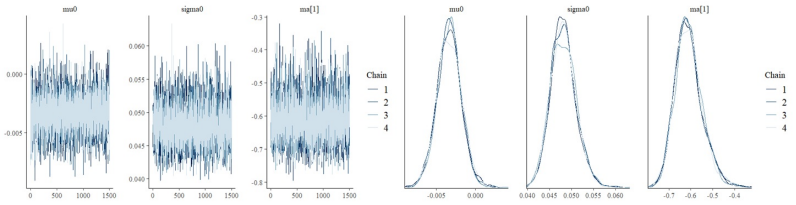


Figure 2.45: Trace and density plots for key Bayesian ARIMA parameters (μ_0, σ_0, and ma[1]) across four chains. Left: trace plots; Right: posterior densities.

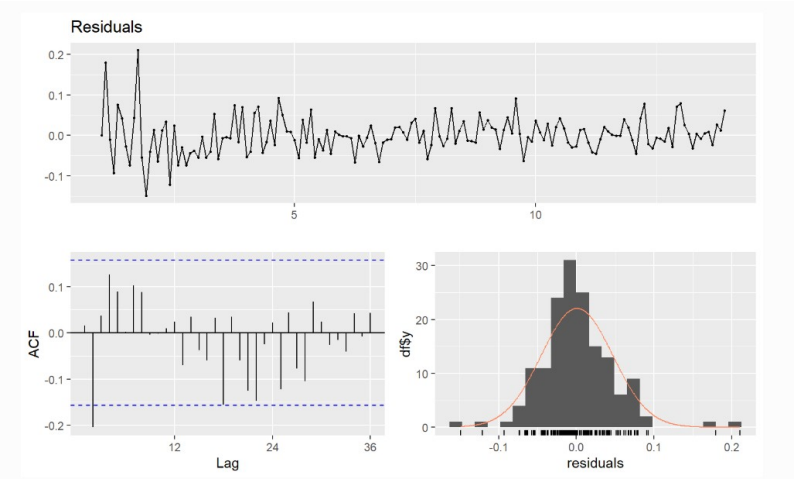


Figure 2.46: Residual diagnostics for the Bayesian ARIMA model: residual time series (top), ACF (bottom-left), and histogram with density overlay (bottom-right).

2.9.3 Predictive Modeling Using Classical Time Series Approaches

To translate the fitted serial structure into prospective statements about prescribing, we generated multistep forecasts from the selected SARIMA specification, ARIMA$(0,1,1)(1,0,0)_{12}$ with drift. Forecasts were produced dynamically from the final observed month, with seasonal recursion capturing annual peaks and troughs and innovation uncertainty propagated to yield fan–shaped 80% and 95% prediction intervals. The point path in Figure 2.47 follows the established post–campaign trajectory—gradual decline on the level scale consistent with the negative drift estimated in differences—while reproducing the amplitude and phase of winter seasonality. As horizon increases, interval width expands at the expected $\sqrt{h}$ rate for integrated processes, yet the forecast band remains well anchored around recent levels, indicating that short–memory MA(1) and seasonal AR(1) components suffice to explain residual persistence without spurious long–range trends [15]. Read alongside the segmented ITS estimates, these classical forecasts offer a transparent counterfactual for stewardship monitoring: near–term predictions accommodate predictable seasonality and short–term shocks, and the accompanying intervals quantify the range of prescribing trajectories consistent with recent dynamics [15, 18].

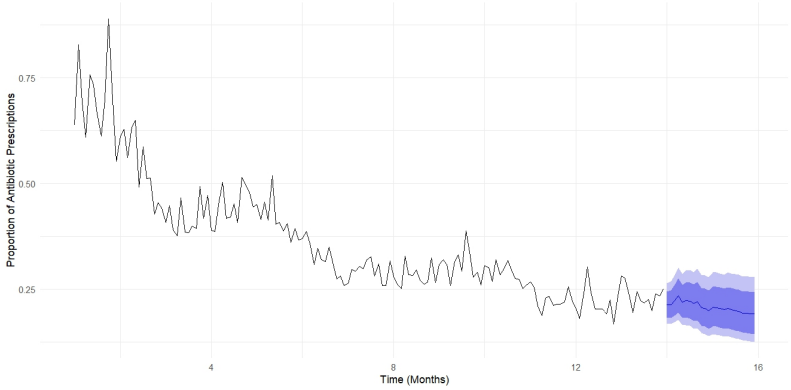


Figure 2.47: Forecasting using classical ARIMA model.

2.9.4 Bayesian Time Series Forecasting with Stan

Using the posterior from the ARIMA$(0,1,1)(1,0,0)_{12}$ model with drift, multi–step forecasts were generated by simulating future innovations jointly with draws of the MA(1), seasonal AR(1), drift, and innovation scale from their posterior, yielding a full posterior predictive distribution at each horizon. The forecast median follows the same gradual post–campaign decline seen in the classical fit while reproducing winter seasonality through the seasonal autoregression; credible intervals widen with horizon as uncertainty accumulates in both states and parameters (Figure 2.48). Relative to the classical fan, Bayesian bands are modestly broader in the near term—reflecting parameter uncertainty—yet remain well anchored around recent levels because the strongly negative MA(1) and modestly negative seasonal AR(1) terms constrain short–memory and annual dynamics already identified in estimation. This alignment between posterior predictive paths and the earlier frequentist trajectory supports the robustness of the inferred decline, while the Bayesian framing provides probability–calibrated uncertainty that integrates

estimation and process noise in a single output suitable for stewardship monitoring and scenario analysis [15, 46, 49].

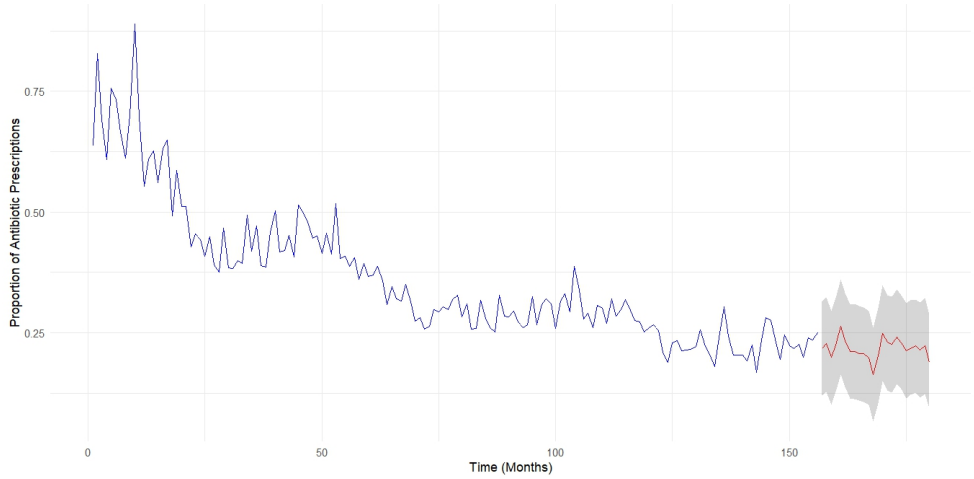


Figure 2.48: Forecasting using Bayesian ARIMA model.

2.9.5 Bayesian Structural Time Series (BSTS)

To complement ARIMA–based inference with an explicitly component–wise view, we estimated a Bayesian Structural Time Series model comprising a local–linear trend, a seasonal state at period 12, and a regression component for the stewardship intervention. This state–space formulation yields posterior distributions for each latent component and for the intervention effect, so that trends, seasonality, and policy signals can be read separately and then recombined for forecasting. Computation proceeded with Hamiltonian Monte Carlo, and convergence diagnostics (rank–normalized $\hat{R}$ and effective sample sizes) were satisfactory, providing a reliable basis for posterior summaries and predictive simulation [46, 49].

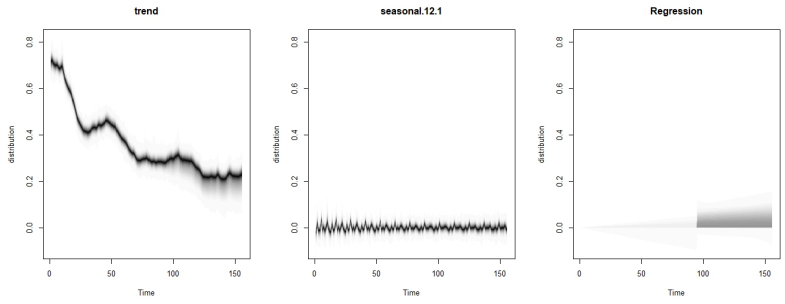


Figure 2.49: Posterior distributions of the trend (left), seasonal (middle), and regression (right) components estimated from the BSTS model.

Posterior marginal distributions for the three components are shown in Figure 2.49. The trend posterior centers below zero on the differenced scale, consistent with the gradual

decline observed in the ARIMA fits; the seasonal component exhibits substantial mass away from zero, reinforcing strong annual structure in prescribing; and the regression (intervention) component concentrates on negative values, indicating a reduction in the post–campaign level after accounting for trend and seasonality. The separation of components clarifies interpretation: the intervention effect is not a surrogate for secular change or seasonal peaks but an additional signal that remains after those dynamics are modeled [15, 18].

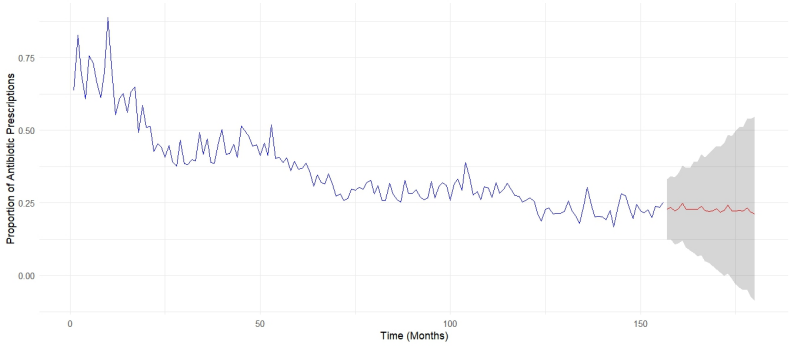


Figure 2.50: BSTS Model Forecast for Antibiotic Prescriptions.

Forecasts drawn from the BSTS posterior predictive distribution (Figure 2.50) track the observed series closely and reproduce winter amplification. Median trajectories align with the classical and Bayesian ARIMA forecasts, while credible intervals are modestly wider in the near term due to explicit integration over latent–state uncertainty. This agreement across modeling frameworks—segmented OLS, SARIMA, Bayesian ARIMA, and BSTS—supports a coherent narrative: inappropriate prescribing declined gradually over the study horizon, with a negative post–intervention shift that persists after controlling for trend and seasonality. The BSTS perspective adds transparent component–level attribution and probability–calibrated forecasts suitable for surveillance and stewardship planning [15, 46].

2.10 Discussion

We applied GLMs to the All of Us cohort to quantify 180–day infection risk after recent clinical procedures and then placed those estimates in dialogue with penalized, Bayesian, machine–learning, and time–series analyses designed for surveillance and policy evaluation. Across frameworks, a consistent clinical picture emerged: respiratory and endocrine disorders, cancer, and invasive procedures—especially intubation and organ transplant—were the most persistent markers of elevated risk, while several demographic factors contributed little after mutual adjustment [47]. This pattern accords with the cohort's construction from linked clinical records and procedures and with prior reporting on the All of Us infrastructure that supports such integrative analyses [30].

In maximum–likelihood GLMs, endocrine disease showed a large and precisely estimated association with infection, whereas the main effect of age was small and uncertain; an age×endocrine interaction did not materially alter that conclusion, a result that is transparent on the odds–ratio scale and grounded in standard GLM theory [47]. Within an

otherwise identical likelihood, Bayesian GLMs clarified how prior information shapes inference when data are modestly informative: weakly informative priors yielded estimates close to the frequentist fit, whereas informative priors anchored to external evidence shifted the age effect while modestly attenuating the endocrine coefficient, with posterior intervals making these movements explicit [36, 49]. The upshot is coherent: the endocrine signal is robust to prior choice, while age is prior–sensitive—precisely the behavior expected when the likelihood carries limited leverage on that parameter [49].

Ridge and LASSO stabilized estimation in the presence of collinearity among comorbidities and procedures, with post–selection (debiased) inference restoring interpretability on the odds–ratio scale [23, 55]. After debiasing, both approaches retained elevated associations for circulatory, respiratory, and endocrine disorders, cancer, and the procedural indicators for intubation and transplant; the multi–comorbidity indicator trended below one after adjustment, consistent with coding and covariate structure. Bayesian penalized analogues with Gaussian (Ridge) and Laplace (LASSO) shrinkage reproduced the same substantive ranking, now with full posterior uncertainty [36]. The concordance between frequentist and Bayesian shrinkage strengthens confidence that the principal signals are not artifacts of a particular regularization scheme [55].

When parsimony was emphasized, stepwise selection (AIC) delivered a compact specification that retained clinically plausible predictors—including respiratory, endocrine, and cancer disorders and the procedure indicators—while discarding weak contributors [55]. In parallel, Bayesian model comparison reframed the exercise around out–of–sample fit: leave–one–out cross–validation favored moderately complex models without overfitting, and projection–predictive selection showed that small submodels preserved accuracy and AUC close to a well–predictive reference, clarifying which variables are essential for generalization [46, 63]. Together, these analyses showed that the same core risk factors are repeatedly prioritized across distinct selection principles.

Probabilistic classifiers provided complementary perspectives focused on prediction. Naive Bayes offered competitive discrimination with simple calibration, and a Bayesian smoothing of the same model traded a small increase in Brier score for higher recall—exactly the regularization behavior anticipated under Laplace smoothing [55, 59]. Bayesian networks improved calibration and discrimination when structure learning accommodated clinically plausible dependencies among comorbidities and procedures, aligning qualitative graph structure with operating characteristics [55]. BART achieved strong classification with stable variable importance that again elevated endocrine, respiratory, and procedure indicators, reinforcing the signals identified by GLMs while maintaining probability–calibrated predictions [55, 63].

Moving from independent observations to clustered or longitudinal settings, mixed–effects models, marginal GEEs, and Bayesian hierarchical models converged on the same substantive story while clarifying the scale of inference. Random–intercept GLMMs retained positive fixed effects for endocrine and respiratory disease and for intubation and transplant, with cluster variability summarized via variance components and intraclass correlation [47]. GEEs provided population–averaged effects with robust (sandwich) variance; comparisons of robust versus model–based standard errors indicated that inferences about the main risk factors were not artifacts of working–correlation choices [65, 67]. Bayesian hierarchical models produced coherent posteriors for fixed and random components with convergence assessed via rank–normalized $\hat{R}$ and effective sample sizes, confirming non-

trivial clustering without overturning fixed–effect conclusions [46, 49].

Residual checks and targeted overdispersion analyses provided guardrails for inference, pointing to quasi–likelihood or zero–inflated formulations where appropriate for binary or count outcomes; these adjustments stabilized standard errors without altering the qualitative ranking of key predictors [47, 48]. Multiple imputation (MICE) and fully Bayesian joint modeling (FBJM) offered two principled routes for handling missingness. Pooled GLM estimates after MICE under MCAR/MAR/MNAR scenarios and posterior summaries from Bayesian GLMs fit to imputed data were broadly compatible with FBJM results, indicating that conclusions are robust to realistic missing–data mechanisms when imputation models are properly specified [24, 26, 27, 69].

We also implemented interrupted time-series models on the monthly prescribing data, which complemented the cohort findings by quantifying intervention impacts at the population level. For monthly antibiotic prescribing following viral diagnoses, classical seasonal ARIMA provided an adequate generative description—first–difference stationarity with short–memory MA(1) and a seasonal AR(1) at period 12—supported by information criteria and white–noise diagnostics [15]. A Bayesian ARIMA with weakly informative priors yielded posterior means close to the frequentist estimates and well–mixed chains, delivering posterior predictive distributions that integrate parameter and disturbance uncertainty; read alongside segmented regression, these fits anchored the stewardship interpretation in counterfactual trajectories with explicit uncertainty [19, 49]. The ITS framing and its time–series counterparts thus support interpretable level and slope contrasts pertinent to stewardship evaluation [17, 18].

Taken together, the chapter demonstrates how inference and prediction for infectious–disease outcomes can be triangulated: interpretable odds–ratio models establish stable associations for a small set of comorbidities and procedures; shrinkage and principled selection guard against overfitting; probabilistic learners validate predictive usefulness and calibration; hierarchical and marginal approaches respect correlation structures; and time–series models extend the lens from individual risk to system behavior with transparent counterfactuals [47, 63]. The consistency of findings across paradigms increases confidence that the identified risk factors reflect the underlying data–generating process, while the few prior–sensitive parameters (e.g., age) mark precisely where external evidence and design knowledge are most informative [27, 36]. This integrated workflow—estimation, shrinkage, selection, predictive validation, correlation–aware modeling, and ITS—offers a reproducible template for monitoring patient–reported outcomes using linked clinical, digital, and social data [15, 16].

Chapter 3

Quantifying Regional Variability in Neural Power Spectra: Stability Mapping and Bayesian Multilevel Modeling

3.1 Introduction

The statistical analysis of long-duration neural signals has expanded rapidly, driven by advances in scalable inference and the recognition that subtle temporal structures encode physiology. Among neural signals, electroencephalography (EEG, i.e., scalp EEG) is particularly rich for quantitative modeling, offering high temporal resolution with multivariate spatial sampling. EEG signals are volume conducted and attenuated by the scalp, skull, and CSF, which act as spatial and temporal filters on the signals. From a statistical standpoint, however, the raw EEG voltage time-series is not the primary object of inference; rather, it is the starting point for a data-reduction and modeling pipeline that preserves physiological information while maintaining interpretability.

Conventional EEG records voltage fluctuations at the scalp, offering excellent temporal resolution but limited spatial specificity because signals are volume conducted through the skull and soft tissue [71, 72]. In contrast, *intracranial* EEG (iEEG) acquires potentials directly from cortical or deep structures [73, 74]. This proximity yields higher signal-to-noise ratios, a broader usable bandwidth, and higher spatial resolution, but it comes at the cost of lower spatial coverage [74–76]. Clinically driven implantation, however, makes iEEG inherently focal and subject specific [73]. Statistically, this entails strong local dependence among neighboring contacts, heterogeneous anatomical coverage across subjects, and heavier-tailed feature distributions due to transient pathological events. In this study, we analyze 16 subjects' long-duration iEEG recordings with a multi-stage analysis designed to address key statistical challenges: structuring the data in a principled way, extracting informative features, and deploying models that capture both the complexity and uncertainty of the signals.

3.1.1 Data Structuring and Windowing

Each continuous iEEG recording is segmented into smaller windows and organized into an experimental design that facilitates inference. In particular, we introduce a two-factor design by extracting comparable recording segments at two times of the day (morning and evening) to incorporate diurnal effects as a fixed factor. Within each segment, we further partition the data into short, non-overlapping time windows (on the order of seconds) that serve as discrete observational units. Treating the time series in this nested, blocked manner (windows within segments) enables us to construct repeated-measures models in which each window serves as a within-subject observation. Rather than averaging the signal over large periods or trials (as is common in event-related potential analyzes), we retain the full window-level distribution of feature estimates to quantify variability directly. This strategy avoids obscuring temporal variability and aligns with contemporary recommendations for robust statistical practice in biometrics [77]. This approach provides a foundation for modeling that explicitly accounts for variation across time and channels rather than treating such variation as noise.

Our cohort comprises 16 subjects, each contributing between 6 and 14 monitoring days (median 9.5 days, inter-quartile range 7–12 days), for a total of 157 days of continuous iEEG recordings. From each day, we extract two non-overlapping 5-minute segments (AM and PM) and subdivide each segment into $J = 15$ non-overlapping 20-s windows, so that an individual subject contributes between 180 and 420 windows per channel. Channel counts per subject range from 110 to 127 bipolar channels (median 122), yielding a total of 4,710 windows across all subjects and, after accounting for all channels, about 598,170 window–channel observations. This variation in per-subject data volume ensures that the multilevel framework can borrow strength across the group while flexibly accommodating individual differences in recording duration and channel coverage.

3.1.2 Spectral Feature Extraction

In a second analytical layer, we switch the focus from the time domain to the frequency domain, since iEEG naturally exhibits oscillatory rhythms (e.g., α, β, γ) that are most directly quantified as spectral peaks. For each window, we estimate the power spectral density (PSD) over the 1–100 Hz band using Welch's averaged periodogram approach with 1 second bins and Hamming tapering [78]. Welch's method (which averages overlapped windowed periodograms) reduces the variance of the raw periodogram without introducing excessive bias, and it remains robust for short, mildly non-stationary segments [79–82]. The resulting window level PSD vectors still span roughly 100 frequency bins (1–100 Hz at 1 Hz resolution), so we apply a parametric spectral decomposition to compress the data into a small set of interpretable parameters. Specifically, we apply the FOOOF algorithm **(short for Fitting Oscillations and One-Over-F)** [7]. FOOOF decomposes each log-PSD into a smooth aperiodic component (capturing the $1/f^{\mathcal{B}}$-like background) and a set of discrete periodic components (narrow-band peaks reflecting oscillatory activity). This decomposition yields a handful of features per window: the aperiodic offset and exponent (which describe the background log-power level and its frequency scaling, respectively); the characteristics of the most prominent oscillatory peak (its center frequency, peak power (amplitude), and bandwidth); and goodness-of-fit metrics (e.g., R^2 and $RMSE$). The aperiodic parameters are particularly interpretable: the offset represents a baseline

log-power level, and the exponent corresponds to the spectral slope. These features can be directly used as predictors in linear models. The peak frequency and power, being localized measures of oscillatory activity, provide insight into dominant rhythms in each window. Importantly, the FOOOF model imposes constraints that keep the extracted features within plausible ranges (for example, enforcing a non-negative bandwidth). This allows us to choose statistical distributions that naturally fit each feature's data. This feature-based approach reduces data dimensionality without discarding meaningful signal structure [7]. It also lays the groundwork for robust statistical modeling by focusing on features that satisfy standard distributional assumptions for inference [83].

Classical dimension-reduction tools (e.g., principal component analysis, independent component analysis, non-negative matrix factorization) summarize spectra as linear mixtures of frequency bins [84–86]. While effective for compression, these linear components are often difficult to interpret physiologically and tend to mix together the aperiodic and oscillatory aspects of the spectrum. In contrast, FOOOF uses a model-based reduction that preserves a direct mapping between its parameters and specific neuro-physiological constructs. For example, the aperiodic offset and exponent quantify the broadband background activity, whereas the peak frequency, bandwidth, and power describe discrete oscillatory processes [7, 87, 88]. This parametric separation keeps features on natural scales and facilitates choosing an appropriate likelihood model for each feature. In summary, FOOOF reduces data dimensionality without sacrificing interpretability or violating standard modeling assumptions, which makes it well-suited for subsequent hierarchical regression on iEEG spectral features.

3.1.3 Exploratory Variability Analysis

We aim here to stratify brain regions by the stability of FOOOF parameters. We use the coefficient of variation (CV) as the stability measure because it is dimensionless and allows for a fair comparison across channels, regions, and subjects. For each subject, region, channel, and recording block (Day×Segment), we compute a CV from the time windows in that block. We then form a region-level CV per block by taking the median across channels present in that block, and we summarize stability for each subject–region–feature by the median of those block values (lower values indicate greater stability). Because all regions for a subject share the same blocks, we test heterogeneity *between* regions with the Friedman test, a rank-based repeated-measures method [89], and report Kendall's W as an effect size for the strength of differences [90]. Within each subject, we adjust p-values across the five features using the Benjamini–Hochberg procedure [91]. We label a subject–feature as *heterogeneous* when the adjusted $p < 0.05$; as *practically similar* when the adjusted $p \geq 0.05$ and $W \leq 0.05$; otherwise, it is *inconclusive*. Combinations with too few regions or too few complete blocks are not evaluable. We adopt $W \leq 0.05$ as a negligible-concordance cutoff: this is deliberately conservative—stricter than the common "small" boundary around $W \approx 0.10$ for Friedman/Kendall effect sizes—and is consistent with agreement scales that classify $W < 0.20$ as only "slight" agreement [92, 93]. While this hypothesis test tells us whether the regions differ or not; we also applied the median CV values to indicate how the regions differ by providing a rank-order of region stability for each subject. When the Friedman test was significant, we performed pairwise post-hoc comparisons between regions using the paired Wilcoxon signed-rank

test on the blocks shared by each region pair, and controlled multiplicity within subject × feature using Benjamini–Hochberg FDR adjustment [94]. This exploratory layer provides both inference (adjusted tests with Kendall's W) and a practical stability ranking that informs the downstream multilevel models.

Our scientific aim is to examine the stability of brain regions rather than the agreement among individual sensors. Channels within a region act as technical replicates and inevitably pick up micro-anatomical differences and electrode-specific noise; a strict within-region homogeneity test therefore answers a different question and can penalize regions that simply have more channels or minor local variability. In contrast, comparing regions within a subject on the same Day×Segment blocks directly tests the anatomical effect of interest while controlling for day/segment factors. The region-level median CV (across channels within block, then across blocks) attenuates idiosyncratic channels and outlying blocks, and yields a single, interpretable stability score per region. This approach provides power and fairness across subjects with different channel counts, and crucially produces the ordered, subject-specific scale required for stratification even when channels inside a region are not perfectly concordant.

3.1.4 Hierarchical Modeling Strategy

To formally test our primary hypotheses – specifically, whether the two target regions of interest (amygdala and hippocampus) differ from the reference region (white matter) after accounting for diurnal and time effects, we employ a Bayesian hierarchical modeling framework [95]. Although white matter has traditionally been assumed to lack narrow-band rhythmic peaks, recent intracranial recordings demonstrate that white matter depth contacts can exhibit informative oscillatory power and phase dynamics [8, 96]. We therefore adopt white matter as a conservative reference while acknowledging its potential physiological contributions. A multilevel (hierarchical) model is well suited to these data given their nested structure: windows are nested within segments, segments within channels, and channels within subjects [83]. We explicitly model this hierarchy by including random intercepts for each channel and for each subject. This allows the model to absorb subject-specific and channel-specific baseline shifts. For fixed (population-level) effects, we include terms that capture our study design factors and temporal trends: (a) a binary indicator for Segment (morning vs. evening) to model diurnal differences; (b) a second-degree polynomial in Time (within each 5-minute segment) to capture within-segment temporal trends (allowing for a convex trajectory over time); this choice has a physiological basis: slow, minute-scale drifts in spectral power can reflect neural adaptation, changes in vigilance or arousal, and electrode–tissue interface shifts during prolonged recordings [97, 98] and (c) Region indicators for the amygdala and hippocampus (with white matter as the reference category).

Figure 3.1 makes clear that the amygdala and hippocampus are gray-matter nuclei embedded in and interconnected by limbic white-matter tracts such as the fornix. They are interconnected not only via the fornix but also by the stria terminalis and angular bundle—pathways shown in both human tractography and animal tracer studies [99, 100]. White matter is predominantly axonal and typically exhibits lower broadband power and fewer narrow band oscillatory peaks than adjacent gray matter [72]. Using white matter as the reference, therefore, provides a conservative references when testing the amygdala

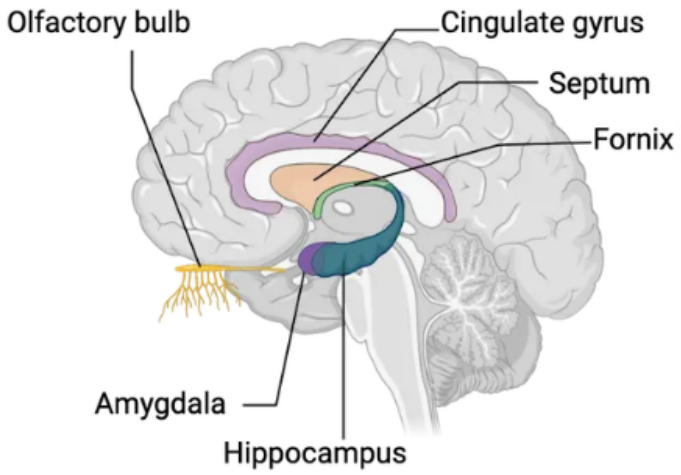


Figure 3.1: Simplified limbic-system schematic highlighting the amygdala (teal) and hippocampus (blue) together with a major white-matter pathway (the fornix) and adjacent cortical regions. Adapted from OpenStax, *Introduction to Behavioral Neuroscience*, Fig. 1.31 (CC BY 4.0) [1].

vs. white matter and the hippocampus vs. white matter; any detected differences are more plausibly attributable to region-specific neural generators rather than global recording or instrumentation effects.

We include interaction terms where needed — for example, Segment × Time interactions to allow the temporal trends to differ between morning and evening segments. This fixed-effect structure encodes our experimental design and hypotheses: it adjusts for circadian phase and within-segment time effects while isolating differences between the brain regions of interest. Crucially, by modeling at the window level, we leverage the full dataset of thousands of observations. This greatly increases statistical power and enables us to estimate variability at multiple levels (window, channel, subject).

Choosing appropriate likelihood functions for each spectral feature is critical to our modeling strategy. Rather than assuming all outcomes are normally distributed, we assign each feature a likelihood distribution that reflects its empirical behavior [101]. For example, the aperiodic offset and exponent features have roughly symmetric distributions but with heavy tails (i.e., a few windows yield unusually extreme values).The dominant oscillation's peak frequency is strictly positive and right-skewed (many observations cluster at lower frequencies, with a long tail to higher frequencies). Similarly, the peak power (amplitude) is positive and right-skewed. By aligning each outcome with an appropriate likelihood, we ensure that our statistical assumptions are congruent with the data's properties. This congruence helps guard against model misspecification and can improve predictive performance.

We use weakly informative priors to stabilize estimation without imposing strong assumptions. For example, each regression coefficient β_j receives a weak Normal prior centered at zero $\beta_j \sim \mathcal{N}(0, 1000)$. This choice is effectively flat on standardized predictors, providing minimal regularization while still preventing the sampler from wandering into

implausible extremes [102]. Although more concentrated priors such as $\mathcal{N}(0, 2.5)$ are often recommended for standardized effects [103], we retained the default specification to maintain consistency across models. We place half-Student-t priors with low degrees of freedom on scale parameters (e.g., residual and random-effect standard deviations) as a practical alternative to a half-Cauchy. Any correlation matrices (if present) receive a Lewandowski-Kurowicka-Joe (LKJ) prior [104] with $\eta = 1$. (Full model specifications and sensitivity analyzes are provided in Section 2.4.2.). We fit the Bayesian models using Hamiltonian Monte Carlo sampling (No-U-Turn Sampler, via Stan's NUTS algorithm), with the *brms* interface [95]. The Bayesian framework provides a cohesive approach to uncertainty. Instead of yielding point estimates alone, it produces a posterior distribution for each parameter (and each derived quantity). We summarize these distributions using credible intervals to indicate the range of plausible values. We validate the fitted models using extensive posterior predictive checks and information criteria.

In posterior predictive checks (PPCs), we simulate data from the fitted model and compare these simulations to the observed data on various statistics (e.g., means, variances, distributional shapes). This allows us to visually and quantitatively assess whether the model can reproduce key features of the real dataset [105]. If the model fits well, the simulated data should closely resemble the observed data on these diagnostics. Systematic discrepancies indicate model inadequacy – for example, if the model consistently underestimates the variability in certain channels or overestimates the power in a particular frequency band. We also evaluate model fit using information criteria. In particular, we compute the Widely Applicable Information Criterion (WAIC) and perform leave-one-out cross-validation (LOO) to estimate out-of-sample predictive accuracy [106]. Here, LOO systematically omits one window–channel observation at a time (while retaining the full predictor set) to assess how well the model generalizes to unseen data points. These criteria penalize model complexity, thereby helping to guard against overfitting, and they guide us in comparing alternative models when needed. Finally, we monitor the fitted degrees-of-freedom parameter in the Student–t likelihoods to assess tail heaviness: very small ν estimates (say, $\nu \lesssim 4$) flag the possible presence of outliers or other influential observations.

In summary, this analysis serves as a case study in combining feature extraction with hierarchical modeling to address the statistical challenges of long-duration neural signals. Key contributions include:

- Structured Windowing: We segment continuous iEEG into a balanced set of short windows (capturing time-of-day and temporal-order factors), yielding a hierarchical longitudinal dataset suitable for multilevel analysis. This bridges raw time series and hierarchical modeling.

- FOOOF-based dimensionality reduction: By modeling each window's spectrum with FOOOF, we reduce the data to a small set of interpretable spectral features. This feature-focused reduction improves interpretability and keeps the analysis grounded in neuro-physiologically meaningful parameters.

- Integration of descriptive & inferential steps: We use the coefficient of variation (CV) as a scale-free stability metric and compute a robust region summary—the median CV across channels within each block, followed by the median across

blocks—to obtain a subject-specific ranking. We then test between-region differences with Friedman's rank-based repeated-measures test (blocks as Day×Segment), adjust p-values across features within subjects using the Benjamini–Hochberg procedure, and report Kendall's W as an effect size. The ranking and tests together support regional stratification and guide the Bayesian multilevel models (e.g., including `Region` and allowing effects to vary by region).

- Rigorous model validation: We perform extensive model checks using posterior predictive simulations and information criteria, following current best practices in Bayesian data analysis. This emphasis on thorough validation helps ensure that our findings are reliable and not artifacts of model misfit [105, 106].

Although our application focuses on iEEG, the statistical principles and strategies we employ — variance partitioning across levels, feature extraction for high-dimensional signals, hierarchical modeling of nested data, and thorough model validation — are broadly applicable to other complex neural signal datasets. Overall, this approach highlights the value of careful statistical design in extracting meaningful insights from long, multivariate time-series data.

3.2 Method

3.2.1 Data Specification

Consider the set of long-duration iEEG recordings from $N = 16$ subjects. We index subjects by $p = 1, \ldots, 16$, and let D_p be the number of monitoring days for subject p. In this cohort, the day counts are $D = (11, 14, 9, 6, 13, 9, 12, 8, 7, 10, 10, 9, 12, 11, 10, 6)$ for subjects 1 through 16, respectively. For each subject's recording, we extracted two non-overlapping five-minute segments for analysis, labeled Segment x (11:50-11:55 AM) and Segment y (11:50-11:55 PM), provided that these intervals lay entirely within the recording's duration. (If a recording did not cover one of these intervals, that segment was skipped for that subject.) This selection of two segments provides an explicit diurnal contrast (morning vs. evening) that we include in the statistical models as a fixed factor (Segment). We further partitioned each 5-minute segment into $J = 15$ contiguous, non-overlapping time windows, denoted $T_1, T_2, \ldots, T_{15}$, where each T_j is a 20-second window of iEEG data. Each day contributes $2 \cdot J = 30$ windows (15 from the AM segment and 15 from the PM segment), so the total number of windows for subject p is $W_p = 2 \cdot J \cdot D_p = 30 \cdot D_p$.

For example, the window counts per subject are

$$(W_1, \ldots, W_{16}) = \begin{matrix}(330,\ 420,\ 270,\ 180,\ 390,\ 270,\ 360,\ 240,\\ 210,\ 300,\ 300,\ 270,\ 360,\ 330,\ 300,\ 180),\end{matrix}$$

corresponding to the D_p values above. In total, the dataset comprises

$$\sum_{p=1}^{16} W_p = 4{,}710$$

non-overlapping 20-second iEEG windows, which serve as the input for spectral feature extraction and subsequent hierarchical analysis.

The iEEG was recorded at a sampling rate of 2048 Hz. Each EDF header provides, for channel c of subject p, a scaling factor $\gamma_{p,c}$ (counts $\rightarrow \mu$V) and a DC offset $\delta_{p,c}$ (counts). Thus, each integer sample $r_{p,c}(t)$ is converted to physical voltage by

$$V_{p,c}(t) \;=\; \frac{r_{p,c}(t) - \delta_{p,c}}{\gamma_{p,c}} \quad (\mu\mathrm{V}).$$

Line-noise attenuation: We then pass each calibrated trace $V_{p,c}(t)$ through a fourth-order digital notch filter (band-stop) targeting 59-61 Hz. The filter is applied in both forward and reverse directions to achieve zero phase shift. Denoting this operation as $\mathcal{N}_{59:61}[\cdot]$, the filtered signal is

$$U_{p,c}(t) = \mathcal{N}_{59:61}\big[V_{p,c}(t)\big].$$

In effect, this zero-phase IIR notch filter suppresses 60 Hz line noise while preserving the waveform's timing.

Bipolar re-referencing: We re-reference the data to bipolar channels by differencing adjacent monopolar leads. If there are 128 monopolar contacts per subject, this yields $C_{\mathrm{bi}} = 127$ bipolar channels. In notation:

$$B_{p,k}(t) = U_{p,k}(t) - U_{p,k+1}(t), \qquad k = 1, \ldots, 127,$$

for each electrode shaft individually, each contact (except the last on each shaft) yields one bipolar derivation. This approach enhances local spatial specificity and reduces line noise and common-mode artifacts; however, it can also attenuate globally coherent low-frequency oscillations by subtracting shared signals between neighboring contacts.

Windowing scheme: We formalize the window indexing as follows. For subject p, channel k, segment $s \in \{x, y\}$, and window index $j \in 1, \ldots, 15$, define the windowed signal

$$B_{p,k,d,s,j}(t) = B_{p,k,d}\big(t_0 + s \cdot \Delta_{\mathrm{seg}} + (j-1) \cdot \Delta_{\mathrm{win}} + t\big), \quad t \in [0, \Delta_{\mathrm{win}}),$$

where t_0 is the start time of the first (AM) segment, $\Delta_{\mathrm{seg}} = 5$ min is the segment length, and $\Delta_{\mathrm{win}} = 20$ s is the window length. In words, $B_{p,k,d,s,j}(t)$ represents the time series for subject p, channel k, day d, during segment s ($s = 0$ for the AM segment, $s = 1$ for the PM segment), in window j. It is a 20-second snippet of the bipolar signal starting at time $(t_0 + s \cdot \Delta_{\mathrm{seg}} + (j-1) \cdot \Delta_{\mathrm{win}})$.

Resulting data structure. Following this schema, each data fragment can be indexed by a 5-tuple (p, d, s, j, k), with:

- $p \in \{1, \ldots, 16\}$: subject identifier,
- $d \in \{1, \ldots, D_p\}$: day index for that subject ($d = 1, \ldots, D_p$),

- $s \in \{x, y\}$: segment (AM or PM),
- $j \in \{1, \ldots, 15\}$: window index within the segment (20-s window),
- $k \in \{1, \ldots, 127\}$: bipolar channel index.

Across all subjects, the total number of monitoring days is $\sum_{p=1}^{16} D_p = 157$ (using the day counts listed above). Since each day contributes 30 windows, there are

$$\sum_{p=1}^{16} W_p \;=\; \sum_{p=1}^{16} 30 D_p \;=\; 30 \times 157 \;=\; 4{,}710$$

total windows in the dataset. Each of these windows contains $\sim$ 122-127 bipolar channel signals based on subjects, so in total, we have $\approx$ 600,000 window–channel observations. These nearly six hundred thousand window–channel data points are the basic units that enter the next steps of spectral feature extraction and hierarchical modeling.

3.2.2 Spectral Estimation

From Fourier Series to a Practical Power Spectrum

Any finite, discrete-time signal $x[n]$ of length N can be written as a weighted sum of orthogonal sinusoids via the *Discrete Fourier Transform* (DFT)

$$X[k] \;=\; \sum_{n=0}^{N-1} x[n]\; e^{-i2\pi kn/N}, \qquad k = 0, \ldots, N-1.$$

The squared magnitude of the DFT,

$$P_{\text{raw}}[k] \;=\; \frac{1}{N f_s}\, |X[k]|^2,$$

is called the **periodogram**. The (raw) periodogram is an asymptotically unbiased but inconsistent estimator of the true power spectral density (PSD); for finite windows, it has high variance and non-negligible bias.

3.2.3 Welch's Variance Reduction Techniques

Welch's method [79] stabilizes the PSD estimate by splitting the signal into *overlapping* segments and averaging their periodograms. The procedure is as follows:

1. **Segment length and resolution:** To obtain a desired frequency resolution of $\Delta f = 1$ Hz with sampling rate $f_s = 2048$ Hz, we choose segment length of $L = \frac{f_s}{\Delta f} = 2048$ samples (≈ 1 s). This yields a frequency grid $f_\ell = \ell$ Hz, for $\ell = 1, \ldots, 100$ over 1-100 Hz band. However, these 1 s Hamming-tapered segments are 50% overlapped and averaged across a 20 s analysis window. Thus each 20 s epoch spans $\approx$ 20 cycles at 1 Hz, ensuring that the slowest rhythms are based on multiple periods even while maintaining 1 Hz spectral resolution and reduced variance.

2. **Taper each segment:** Multiply each segment $x_m[n]$ (of length L) by a *Hamming* window $w[n]$ to mitigate edge discontinuities (reducing spectral leakage). Denote the windowed segment as

$$x_m^{(\text{win})}[n] = w[n]\, x_m[n], \qquad n = 0, \ldots, L-1.$$

3. **50% overlap:** Use overlapping segments offset by $L/2$ samples (50% overlap). Thus each new segment shares half its data with the previous one. This roughly doubles the number of segments M while still providing enough independence to further reduce variance.

4. **Average the periodograms:** Compute the periodogram for each tapered segment and average them. For M segments, the Welch PSD estimate is

$$\widehat{S}[f_\ell] \;=\; \tfrac{1}{M} \sum_{m=1}^{M} \frac{1}{L f_s} \left| \sum_{n=0}^{L-1} w[n]\, x_m[n]\, e^{-i2\pi \ell n / L} \right|^2, \qquad \ell = 1, \ldots, 100.$$

   For a 20 s window of data, using $L = 2048$ with 50% overlap yields

$$M \;=\; 1 + \left\lfloor \frac{N - L}{L/2} \right\rfloor \;=\; 39$$

   segments to average.

Balancing Bias and Variance

Welch's method trades a small increase in bias for a large reduction in variance, as summarized below:

- **Reduced variance:** Averaging M spectra lowers the variance of the PSD estimate by approximately a factor of M. In this study, $M \approx 39$, so random fluctuations are suppressed, yielding a much smoother PSD curve even for a 20 s window.
- **Controlled bias:** The Hamming window taper introduces minimal bias. It limits spectral leakage (energy spread beyond each 1 second bin), so sharp spectral features are not overly smeared.
- **Robust to non-stationarity:** Using short 1 s sub segments (with overlap) makes the Welch estimate tolerant to the mild non-stationarity within a 20 s iEEG window, while still achieving the desired 1 Hz frequency resolution.

In summary, for each 20 s window (of each channel), we obtain a Welch PSD estimate $\widehat{S}_{pdsjk}(f_\ell)$, $\ell = 1, \ldots, 100$. These 1-100 Hz PSD vectors (one per window-channel) feed directly into the next step, the FOOOF-based spectral decomposition and subsequently into the hierarchical modeling.

Dimension Reduction with FOOOF

The Welch PSD estimate $\widehat{S}_{pdsjk}(f_m)$ describes how signal power is distributed across frequency, but its steep $1/f^{\mathcal{B}}$ decline obscures narrowband oscillations that are often of scientific interest. Following [7], we use the *Fitting Oscillations and One-Over-F* (FOOOF) algorithm to separate this broadband "aperiodic" background from superimposed periodic peaks. The starting point is the base-10 logarithm of the PSD,

$$y_m \;=\; \log_{10}\!\big(\widehat{S}_{pdsjk}(f_m)\big), \qquad f_m \in [1, 100]\ \text{Hz},$$

because the aperiodic background is approximately linear in log–log coordinates, while oscillatory peaks appear as localized deviations on that scale.

Aperiodic Background

FOOOF assumes that the log-transformed background follows a power law

$$b(f;\, O, \mathcal{B}) = O - \mathcal{B}\, \log_{10} f, \tag{3.1}$$

where O is the *offset* (intercept) and $\mathcal{B}$ is the *exponent* (slope). Equation 3.1 is fitted by ordinary least squares to the smoothed log-spectrum; the residuals $e_m = y_m - b(f_m; O, \mathcal{B})$ highlight candidate oscillatory peaks.

Peak Detection

Residuals are scanned for local maxima that exceed a data-driven threshold (set to two standard deviations of e_m by default). Each candidate peak centered at frequency μ_q is modeled as a Gaussian in linear frequency (on the log-PSD scale):

$$g_q(f;\, \mu_q, \sigma_q, A_q) = A_q\, \exp\!\Big[-\frac{(f-\mu_q)^2}{2\sigma_q^2}\Big],$$

parameterised by centre frequency μ_q (Hz), bandwidth σ_q (Hz), and amplitude A_q (log-power units). Peaks are fit jointly with constraints $\mu_q \in [1, 100]$ Hz and $\sigma_q > 0$. The complete spectral model is

$$\widehat{y}_m \;=\; b(f_m; O, \mathcal{B}) \;+\; \sum_{q=1}^{Q} g_q(f_m;\, \mu_q, \sigma_q, A_q),$$

where Q is the number of detected peaks. Goodness of fit is summarized by the coefficient of determination

$$R^2 = 1 - \frac{\sum_m (y_m - \widehat{y}_m)^2}{\sum_m (y_m - \bar{y})^2}, \qquad \bar{y} = \frac{1}{100}\sum_m y_m,$$

and the root-mean-square error $\mathrm{RMSE} = \big[\frac{1}{100}\sum_m (y_m - \widehat{y}_m)^2\big]^{1/2}$.

For each window, we retain (i) the two aperiodic parameters O and $\mathcal{B}$, (ii) the single peak with the largest amplitude A_q and record its center frequency μ_q, power A_q, and bandwidth $w_q = 2\sqrt{2\ln 2}\,\sigma_q$ (full width at half-maximum), and (iii) the fit diagnostics R^2 and RMSE. Thus, every PSD vector is distilled into a seven-dimensional feature vector

$$\big(O,\ \mathcal{B},\ \mu^*,\ A^*,\ w^*,\ R^2,\ \mathrm{RMSE}\big),$$

where the superscript * denotes the peak with maximum power A_q.

3.2.4 Statistical Analysis

Assessing Regional Stability with the CV

To obtain a data-driven regional stratification, we evaluated whether brain regions differed from one another in the stability of FOOOF parameters computed from channels within each region. Let $p \in \{1, \ldots, 16\}$ index subjects, $r \in \mathcal{R}$ regions, k channels within a region, $j \in \{1, \ldots, J\}$ the time windows per block ($J = 15$), and let the recording block be $b \equiv (d, s)$ for Day $\times$ Segment. For a given feature θ (offset, exponent, center frequency, power, bandwidth), we denoted the window value by θ_{prkbj}.

For each subject–region–channel–block, we computed CV across windows,

$$\bar{\theta}_{prkb} = \frac{1}{J} \sum_{j=1}^{J} \theta_{prkbj},$$

$$s^2_{prkb} = \frac{1}{J-1} \sum_{j=1}^{J} \left(\theta_{prkbj} - \bar{\theta}_{prkb}\right)^2,$$

$$\mathrm{CV}_{prkb}(\theta) = \frac{s_{prkb}}{|\bar{\theta}_{prkb}|}$$

defined when $\bar{\theta}_{prkb} \neq 0$.

Robust region summary: Let $\mathcal{D}_p$ denote the set of Day $\times$ Segment blocks observed for subject p, and let $\mathcal{D}_p^\star \subseteq \mathcal{D}_p$ be the subset of complete blocks (all regions present). We then aggregated across channels to obtain a block-wise region CV,

$$M_{prb}(\theta) = \mathrm{median}_{k \in r}\{\mathrm{CV}_{prkb}(\theta)\},$$

and summarized region stability for subject p by the median across blocks,

$$S_{pr}(\theta) = \mathrm{median}_{b \in \mathcal{D}_p} M_{prb}(\theta).$$

Lower values of $S_{pr}(\theta)$ indicate greater temporal stability.

Between-region heterogeneity (Friedman test): For each subject p and feature θ, we tested the null hypothesis $H_0^{(p,\theta)} : \left\{M_{prb}(\theta) : b \in \mathcal{D}_p^\star\right\}$ had identical distributions across $r \in \mathcal{R} \iff \mathbb{E}[R_{pr}] = \frac{m(k+1)}{2}\ \forall r,$, while the alternative hypothesis is $H_1^{(p,\theta)} : \exists\, r \neq r'$ such that the distributions differ.

Since the regions for a subject were observed on the same blocks, we tested *between-region* differences using the Friedman rank-based repeated-measures procedure [89]. For each (p, θ), we restricted inference to the set of complete blocks $\mathcal{D}_p^\star \subseteq \mathcal{D}_p$ in which all regions under comparison were present; let $m = |\mathcal{D}_p^\star|$ and $k = |\mathcal{R}|$. Within each block, we ranked $\{M_{prb}(\theta) : r \in \mathcal{R}\}$ from most to least stable (smallest to largest CV). Denoting by $R_{pr} = \sum_{b \in \mathcal{D}_p^\star} \mathrm{rank}\{M_{prb}(\theta)\}$ the sum of ranks for region r, the Friedman statistic was

$$Q_p(\theta) = \frac{12}{m\, k(k+1)} \sum_{r \in \mathcal{R}} R_{pr}^2 - 3m(k+1),$$

which is approximately χ^2_{k-1} under the null. As an effect size, we reported Kendall's coefficient of concordance [90],

$$W_p(\theta) = \frac{Q_p(\theta)}{m\,(k-1)} \in [0,1],$$

interpretable as the proportion of rank variance explained by region. Within each subject, we controlled for multiplicity across the five features by using the Benjamini–Hochberg adjustment of the Friedman p-values [91].

False–discovery control within subject (Benjamini-Hochberg): For each subject p we tested the $m = 5$ features $\theta \in \Theta$ with the Friedman procedure and obtained raw p–values $\{\, p_p(\theta) : \theta \in \Theta \,\}$. To control the expected proportion of false rejections among all rejections (the false discovery rate, FDR) at level α, we applied the Benjamini–Hochberg (BH) adjustment [91]. Let $p_{p,(1)} \le \cdots \le p_{p,(m)}$ be the ordered p–values with corresponding features $\theta_{(1)}, \ldots, \theta_{(m)}$. The BH step–up rejection rule is

$$i^\star \;=\; \max\left\{ i \in \{1,\ldots,m\} : p_{p,(i)} \le \alpha\,\frac{i}{m} \right\}, \qquad \text{reject } H_0^{(p,\theta_{(i)})} \text{ for } i = 1,\ldots,i^\star.$$

Equivalently, BH produces monotone adjusted p–values

$$q_{p,(i)} \;=\; \min_{j \ge i}\left\{ \frac{m}{j}\, p_{p,(j)} \right\} \;\wedge\; 1, \qquad \text{and we set } q_p(\theta_{(i)}) = q_{p,(i)}.$$

Using BH is important here because we make one comparison per feature within each subject; adjusting across these five tests controls multiplicity with higher power than family-wise procedures (e.g., Bonferroni) while keeping the FDR at or below α under independence or mild positive dependence among features [91].

Decision rule and post-hoc contrasts: Let $q_p(\theta)$ denote the BH-adjusted Friedman p-value. We used $\alpha = 0.05$ and a small-effect threshold $w_0 = 0.05$. We declared *heterogeneity* if $q_p(\theta) < \alpha$; *practical similarity* if $q_p(\theta) \ge \alpha$ and $W_p(\theta) \le w_0$; otherwise the result was *inconclusive*. Combinations with fewer than two regions ($k < 2$) or fewer than three complete blocks ($m < 3$) were classified as *not evaluable*. When $H_0^{(p,\theta)}$ was rejected, we localized differences with *paired Wilcoxon signed-rank tests* run for each region pair on the *blocks shared by that pair*, and controlled multiplicity *within subject* × *feature* using the Benjamini–Hochberg false-discovery procedure [91, 94]. Independently of significance, we used $S_{pr}(\theta)$ to provide a subject-specific ordering (lower S_{pr} = more stable) for regional stratification.

Hierarchical Modeling

Another goal of our analysis is to quantify how strongly two areas of interest (the amygdala and hippocampus) differ from a reference area (white matter) once diurnal segments and within-segment times have been taken into account. Because observations are nested (channels within subjects) and the residual variability is heavy-tailed for several FOOOF parameters, we adopt a Bayesian hierarchical framework fitted with Hamiltonian Monte Carlo via `brms` and Stan [34, 95].

Data Structure and Notation

Let segment $s = 0$ correspond to the [11:50–11:55] AM window (label x) and $s = 1$ correspond to the [23:50–23:55] PM window (label y). This makes it explicit that s is treated as a numeric dummy variable in formulas (e.g., in 3.2), while x/y are merely labels for exposition. We encode the time of day as a binary indicator.

$$s = \begin{cases} 0, & \text{AM segment,} \\ 1, & \text{PM segment,} \end{cases}$$

which though inherently categorical, is conveniently represented numerically so that s can enter interaction terms with the continuous time polynomial in 3.2. Brain region r is treated as a three-level categorical factor (white matter as baseline, amygdala, hippocampus). Since each observation corresponds to a single 20-s window from one bipolar channel, we index observations by $i = 1, \ldots, N$, and attach the following identifiers and covariates:

p_i	subject $(1, \ldots, 16)$,
k_i	channel (nested within subject),
s_i	segment (0 = AM, 1 = PM),
t_i	time window $(1, \ldots, 15)$,
r_i	cortical region (factor with baseline = white matter).

To compare the two treatment regions with the baseline, we create dummy variables.

$$r_i^{(A)} = \mathbb{1}\{r_i = \text{amygdala}\}, \qquad r_i^{(H)} = \mathbb{1}\{r_i = \text{hippocampus}\},$$

so that $r_i^{(A)} = r_i^{(H)} = 0$ denotes white matter.

Fixed Effects

To capture the approximately convex pattern of all five FOOOF parameters, we considered a second-degree polynomial that ensures curvature with minimal complexity. Since the morning and evening trajectories plateau at different levels, we include $Segment \times Time$ interactions. The full fixed-effects block, therefore, consists of

$$1,\ s_i,\ t_{1i},\ t_{2i},\ s_i t_{1i},\ s_i t_{2i},\ r_i^{(A)},\ r_i^{(H)}.$$

Random Effects

To account for nested sampling, we include random intercepts

$$u_p \sim \mathcal{N}(0, \sigma^2_{\text{subject}}), \qquad v_{pc} \sim \mathcal{N}(0, \sigma^2_{\text{channel}}),$$

where v_{pc} is nested within u_p.

Linear Predictor

Putting the pieces together, the linear predictor for observation i is

$$\eta_i = \beta_0 + \beta_S s_i + \beta_{T1} t_{1i} + \beta_{T2} t_{2i} + \beta_{ST1} s_i t_{1i} + \beta_{ST2} s_i t_{2i} + \beta_A r_i^{(A)} + \beta_H r_i^{(H)} + u_{p_i} + v_{p_i k_i}. \tag{3.2}$$

Likelihood Families

Each FOOOF parameter θ_i is modeled with a distribution whose shape matches its empirical histogram:

- **Offset (O) and Exponent ($\mathcal{B}$):** Student–$t(\nu, \mu = \eta_i, \sigma)$ to accommodate heavy tails [101]. Here μ is the location parameter and is equal to the mean only if $\nu > 1$.
- **Centre frequency (μ^*):** Let z_i denote the center frequency of the dominant peak in window i. Because $z_i > 0$ and is typically right-skewed, we model it with a Gamma generalized linear model and a log link:

$$z_i \sim \text{Gamma}\big(\kappa,\ \text{rate} = \kappa/\mu_i\big), \qquad \log \mu_i = \eta_i.$$

  Here $\kappa > 0$ is the shape parameter, η_i is the linear predictor in (3.2), and $\mu_i = \mathbb{E}(z_i \mid \eta_i) = \exp(\eta_i)$ is the conditional mean. The rate parameter κ/μ_i guaranties $\mathbb{E}(z_i) = \mu_i$ while allowing the variance $\text{Var}(z_i) = \mu_i^2/\kappa$ to adapt through κ [107].
- **Peak power (A^*):** $\log \theta_i \sim \mathcal{N}(\eta_i, \sigma^2)$ (log-normal likelihood).

Prior Distributions

Under Bayes' theorem, the information from the data, y, combines with prior beliefs about the parameter vector θ through the factorization

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{\int p(y \mid \theta)\, p(\theta)\, d\theta}, \tag{3.3}$$

where $p(y \mid \theta)$ is the likelihood and $p(\theta)$ the prior distribution. Although the window-level sample size is large, multilevel models include variance and correlation components that can remain weakly identified. Classical Bernstein–von Mises (BvM) [105, 108] guaranties that posterior normality and asymptotic prior irrelevance applied to fixed, low-dimensional regular parametric models and do not extend uniformly to hierarchical settings, where latent effects grow with n or under mild misspecification. Consequently, we treat priors as part of the model, not as a vanishing nuisance, and adopt weakly informative [102]. For fixed effects, brms places weak Normal priors; intercepts receive a Student-t prior; and group-level (random-effect) standard deviations are assigned half-Student-t priors. For the Gaussian model of spectral power, the residual standard deviation also receives a half-Student-t prior, while in the Gamma model of center frequency, no residual σ is present.

$$\beta_j \sim \mathcal{N}(0, 1000), \qquad \text{Intercept} \sim \text{Student-}t(3, 0, 10), \qquad \sigma \sim \text{Half-Student-}t(3, 0, 10)$$

and in the Gamma model. These choices inject just enough structure to

1. prevent degenerate chains when a variance component approaches zero,
2. avoid funnel pathologies and improve the numerical stability of Hamiltonian Monte Carlo [102, 109],
3. encode minimal domain knowledge (e.g. positivity, scale) without distorting estimates.

Estimation and Model Checking

We draw four Hamiltonian Monte Carlo (HMC) chains of 4 000 iterations each (2 000 warm-up). Convergence is accepted when $\widehat{R} < 1.01$ and effective sample size > 400 per parameter. Model adequacy is assessed by:

1. **Posterior-predictive checks (PPC):** Simulated replicates are compared with observed data on means, variances and quartiles; systematic discrepancies point to mis-specification [105].
2. **Information criteria:** WAIC and leave-one-out cross-validation (LOO) via Pareto-smoothed importance sampling; the model with higher expected log predictive density is preferred [106].
3. **Heavy-tail monitoring:** For Student-t likelihoods, $\nu < 4$ triggers sensitivity analyses using skew-t forms to confirm robustness.

Interpreting Regional Effects

Because white matter is the reference category,

$$\beta_A = \text{mean difference (amygdala – white matter)},$$

$$\beta_H = \text{mean difference (hippocampus – white matter)}.$$

In the Student-t models, these are additive contrasts. In the Gamma and log-normal models, $\exp \beta_A$ and $\exp \beta_H$ are multiplicative ratios: values > 1 indicate higher center frequency or power in the treatment region; values < 1 indicate lower values. Credible intervals that exclude zero (or one on the ratio scale) provide evidence that treatment and control regions differ after adjusting for diurnal context and temporal drift.

Thus, equation 3.2 delivers subject- and channel-adjusted contrasts that integrate anatomical targeting, circadian phase, and within-segment dynamics, while the Bayesian framework provides a full quantification of uncertainty at every level of the analysis.

From Univariate to Multivariate Modeling

The univariate fits (in equation 3.2) treat each aperiodic feature as an independent outcome. However, we are also interested in applying the multivariate model since the empirical inspection reveals a strong association between Offset and Exponent. Ignoring that dependence understates joint uncertainty and precludes inference on their combined behavior. We therefor, we extended the linear predictor in equation 3.2 to a bivariate response,

$$\mathbf{y}_i := \begin{pmatrix} O_i \\ \mathcal{B}_i \end{pmatrix} \mid \boldsymbol{\eta}_i, \boldsymbol{\Sigma}, \nu \sim t_\nu(\boldsymbol{\eta}_i, \boldsymbol{\Sigma}), \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_O^2 & \rho\, \sigma_O \sigma_{\mathcal{B}} \\ \rho\, \sigma_O \sigma_{\mathcal{B}} & \sigma_{\mathcal{B}}^2 \end{pmatrix}, \tag{3.4}$$

where $\boldsymbol{\eta}_i = (\eta_{Oi}, \eta_{\mathcal{B}i})^\top$ shares the same fixed and random effects as before, ν governs tail heaviness, and ρ is the *residual correlation* that captures any unexplained co-fluctuation. Equation 3.4 nests the two univariate t models ($\rho = 0$) and recovers their marginal posteriors, while additionally providing coherent inference on functions of both outcomes

(e.g. $O - \mathcal{B}$ or their ratio). The theoretical advantages of modeling $\rho \neq 0$ are summarized below.

Building on equation (3.4), we compute a bivariate Student–t model that shares the same fixed- and random-effects structure as the univariate analyses but allows the residuals of Offset and Exponent to correlate. Let $\mathbf{y}_i = (O_i, \mathcal{B}_i)^\top$ for window i and let $\boldsymbol{\eta}_i$ be the two-element vector of linear predictors defined in 3.2. The likelihood is

$$\mathbf{y}_i \mid \boldsymbol{\eta}_i, \Sigma, \nu \;\sim\; t_\nu\big(\boldsymbol{\eta}_i,\, \Sigma\big), \qquad \Sigma = \begin{pmatrix} \sigma_O^2 & \rho\, \sigma_O \sigma_{\mathcal{B}} \\ \rho\, \sigma_O \sigma_{\mathcal{B}} & \sigma_{\mathcal{B}}^2 \end{pmatrix},$$

where ν is a shared degrees-of-freedom parameter and ρ captures residual (conditional) correlation between the two aperiodic features. We used weakly informative priors to mirror the univariate models (LKJ(1) for ρ).

Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iq})^\top$ denote the q-variate response for observational unit i with linear predictor $\boldsymbol{\mu}_i = \mathbb{E}(\mathbf{y}_i \mid \text{covariates})$. The *residual vector* is

$$\mathbf{e}_i = \mathbf{y}_i - \boldsymbol{\mu}_i, \qquad i = 1, \ldots, n,$$

and summaries the part of $\mathbf{y}_i$ that remains unexplained after accounting for all fixed and random effects. Under a correctly specified model the residuals satisfy $\mathbb{E}(\mathbf{e}_i) = \mathbf{0}$ and

$$\operatorname{Var}(\mathbf{e}_i) = \boldsymbol{\Sigma}_e = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1q} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1q} & \sigma_{2q} & \cdots & \sigma_q^2 \end{pmatrix}. \tag{3.5}$$

The off-diagonal element σ_{12} measures the *residual covariance* between responses 1 and 2. A scale-free measure is the *residual correlation*

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2}, \quad -1 \leq \rho_{12} \leq 1.$$

A non-zero ρ_{12} indicates that, *after* controlling for the covariates and hierarchical structure, the two outcomes share variability that is not captured by those covariates. This matters for three reasons:

(i) **Statistical validity.** Ignoring residual correlation (by fitting separate univariate models) treats σ_{12} as 0. If $\sigma_{12} \neq 0$, standard errors for contrasts that involve both outcomes—e.g. $y_{i1} - y_{i2}$—are understated, inflating type-I error rates [110].

(ii) **Efficiency.** Multivariate estimation exploits the information in σ_{12}, so each response "borrows strength" from the other. When $|\rho_{12}| > 0$, joint models yield smaller variances for regression coefficients than independent fits [111].

(iii) **Substantive insight.** The sign and magnitude of ρ_{12} help to characterize underlying physiology. In our application, a positive $\rho_{O\mathcal{B}}$ indicates windows with elevated broadband power (Offset) also tend to have steeper spectral slopes (Exponent), suggesting a shared mechanism influencing both aperiodic parameters.

Consequently, the multivariate formulation in equation (3.5) offers both more accurate inference and richer scientific interpretation than treating each feature in isolation [112].

For the multivariate Student-t model of Offset and Exponent, we also used weakly informative priors. Regression coefficients were assigned $\mathcal{N}(0, 1000)$ priors, and intercepts received Student-$t(3, 0, 10)$ priors on the link scale. Group-level (random-effect) standard deviations and residual standard deviations for each outcome were given half-Student-$t(3, 0, 10)$ priors. Each degree-of-freedom parameter ν of the Student-t likelihood was assigned a Exponential$(1/30)$ prior. Because we allowed the residuals of Offset and Exponent to correlate, the residual correlation matrix received an LKJ(1) prior, which is uniform over all valid correlation matrices [104].

Software Workflow

All signal-processing steps including EDF import, notch filtering, bipolar referencing, and Welch-PSD estimation were performed in MATLAB ® R2022b (The MathWorks, Natick, MA) using the open-source FieldTrip toolbox [113]. FOOOF decomposition was then carried out via MATLAB's *Python* interface, ensuring seamless integration of the Python FOOOF module within our Matlab pipeline. FieldTrip's vectorized I/O and high-throughput pipeline comfortably handle $\approx 6 \times 10^5$ window–channel epochs in memory. The subsequent exploratory analysis with FOOOF parameters and following Bayesian multilevel models was fitted in R 4.4.1 [114] via `brms` [102] and the Stan probabilistic engine [34], which supplies the wide range of likelihood families, multivariate responses, and hierarchical structures required for the analysis of this study.

3.3 Results

3.3.1 Illustrative Raw Spectra and FOOOF Decomposition

We begin with a representative day from one subject to visualize the signals that underlie the cohort analyses. For each 20 s window we compute the Welch PSD and then average over bipolar channels within each anatomical region (amygdala, hippocampus, white matter) and within segment (AM vs. PM).

Aperiodic and Periodic Components via FOOOF

Applying FOOOF to each window–channel PSD separates a smooth aperiodic background from periodic peaks (Section 2.3). The two panels in Fig. 3.3 display the region-wise mean aperiodic background, while 3.4 shows the mean periodic component (residual power after background subtraction) for the same subject/day and both segments. These components are exactly the quantities whose parameters enter the subsequent analyses.

Across all windows and channels, we then extract five features per window: the aperiodic offset and exponent, plus the center frequency, power, and bandwidth of the dominant oscillatory peak. These distilled parameters form the basis for our exploratory variability analysis (Section 3.2) and subsequent hierarchical modeling (Section 3.3).

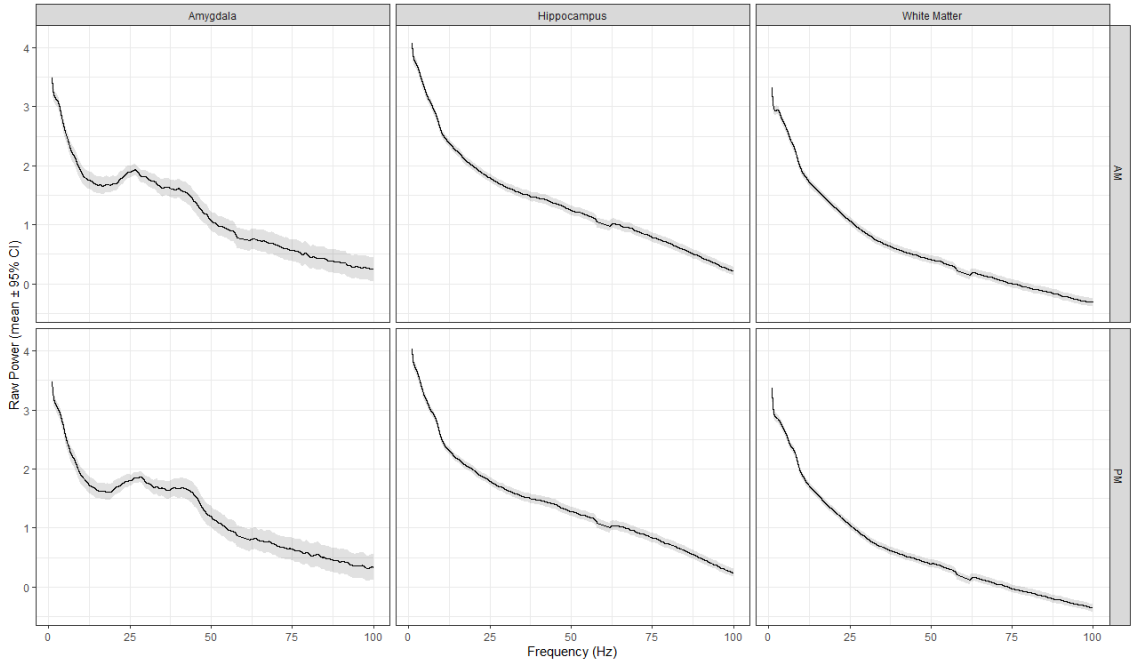


Figure 3.2: Raw power spectral density (1–100 Hz) for a representative subject/day, averaged across channels within each region. Top row: AM; bottom row: PM. The mean PSD (black) and its 95% CI (shaded) show a broadband $1/f$ profile with superimposed narrow peaks, motivating the FOOOF decomposition.

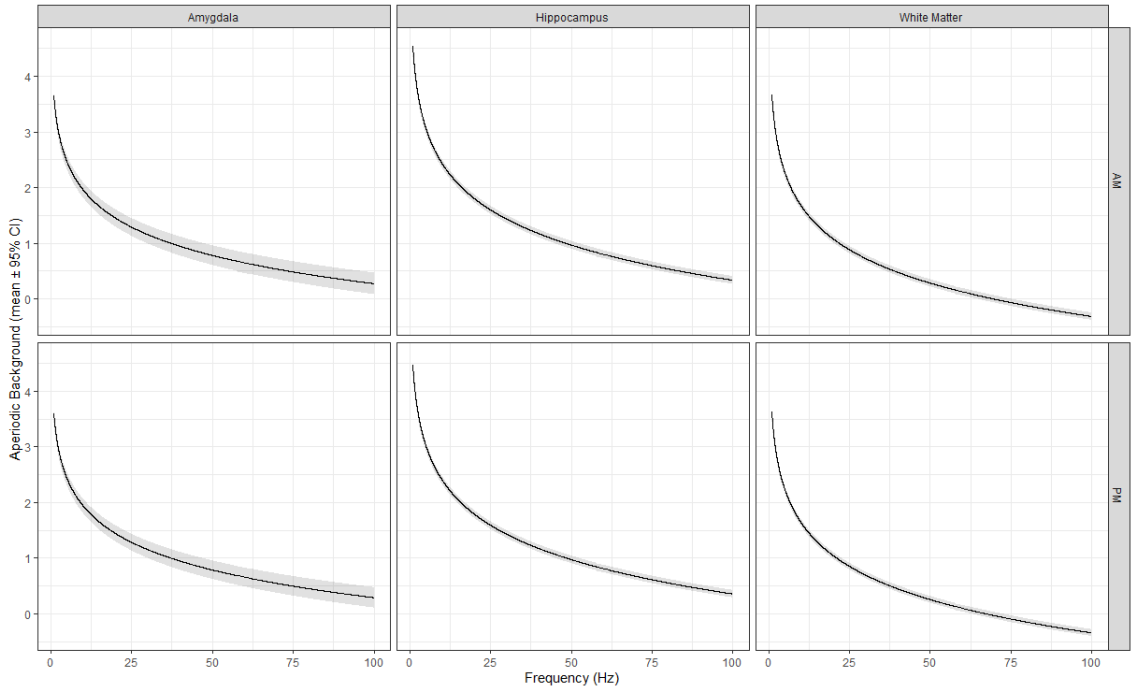


Figure 3.3: Aperiodic background (mean ± 95% CI) by region and segment for the same subject/day. Curves reflect the fitted background component that yields the aperiodic *offset* and *exponent* features.

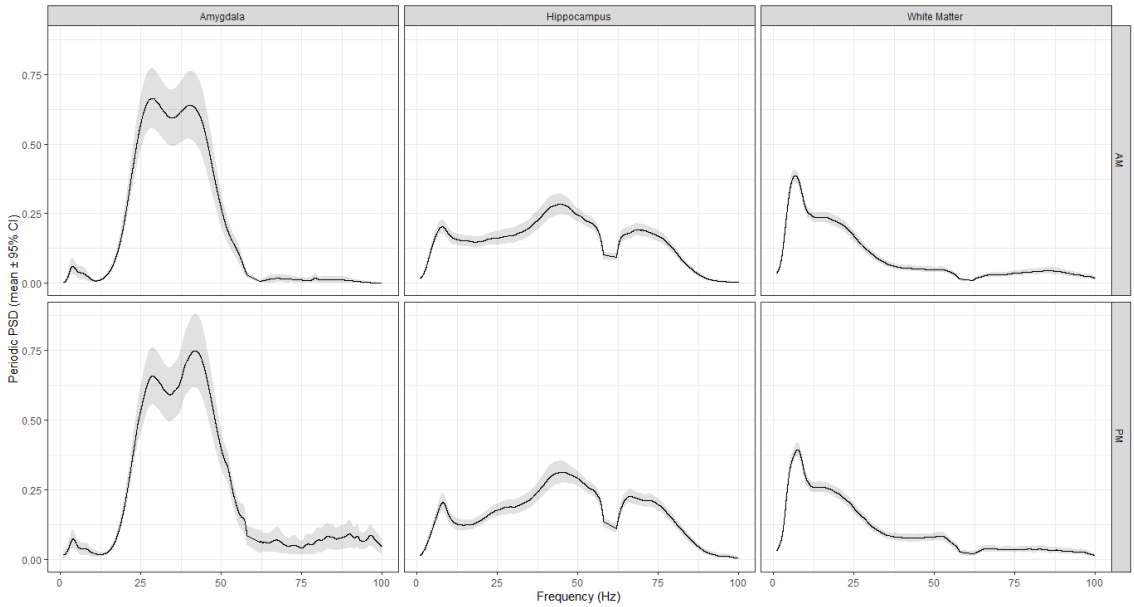


Figure 3.4: Periodic (residual) spectra, i.e., power remaining after subtracting the aperiodic background, by region and segment for the same subject/day. Peaks here give rise to the *center frequency*, *power*, and *bandwidth* features of the dominant oscillation.

Distribution of FOOOF Features

Figure 3.5 illustrates the empirical distributions of the five FOOOF-derived parameters across representative subjects. Each sub-panel (a–e) shows histograms with kernel density overlays, allowing assessment of central tendencies, dispersion, and deviations from normality. These empirical shapes guided the selection of likelihood families in the subsequent hierarchical models.

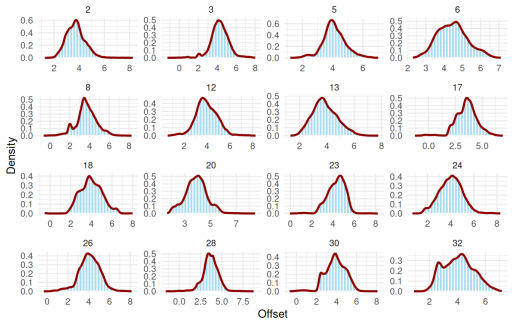
As shown in Fig. 3.5a (panel a), the offset values are approximately symmetric and unimodal for most subjects, with central locations between 3 and 5 on the log-power scale. A small number of subjects exhibit heavier tails or secondary bumps, consistent with occasional outlying windows. This motivates the use of a Student-t likelihood, which can capture both the central bulk and moderate tail heaviness.

The exponent distributions (Fig. 3.5b, panel b) concentrate between 1.5 and 2.5, consistent with the canonical $1/f$ spectral slope. Most subjects show unimodal, near-symmetric forms, though a few display slight skewness or multi-modality, reflecting heterogeneity within recordings. A Student-t family provides a robust model that accommodates these deviations without enforcing strict normality.

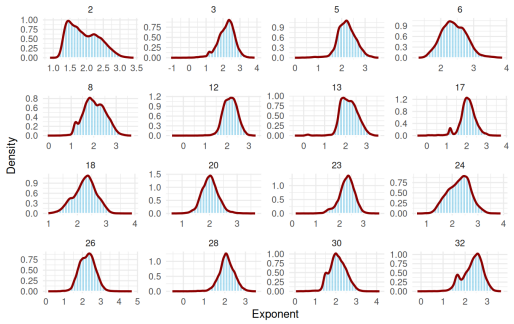
The center frequency distributions (Fig. 3.5c, panel c) are strongly right-skewed, with most peaks located in the 5–20 Hz range, covering θ–α–β rhythms. Several subjects show long tails extending toward higher frequencies, reflecting broader oscillatory activity. The strictly positive, right-skewed form justifies modeling with a Gamma likelihood under a log link.

The power distributions (Fig. 3.5d, panel d) are positive and highly skewed, with sharp concentrations near low amplitudes and a gradual decline toward larger values. This pattern is stable across subjects and aligns well with a log-normal likelihood, which linearized multiplicative variation while keeping interpretation on a natural scale.

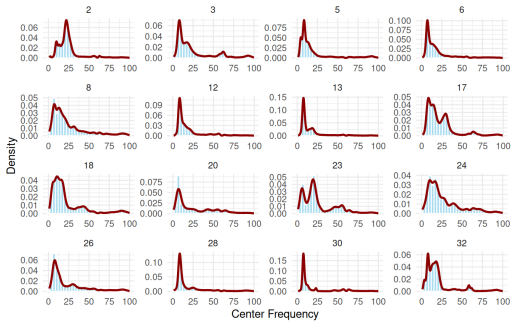
The bandwidth distributions (Fig. 3.5e, panel e) are the most heterogeneous. Many subjects show bimodal or long-tailed shapes, with clusters of narrow peaks but also broad oscillations reaching the upper limit of the analyzed range. Applying a log transform stabilizes the scale, after which a Gaussian approximation is suitable; in subjects with heavier tails, Student-*t* variants provide additional robustness.



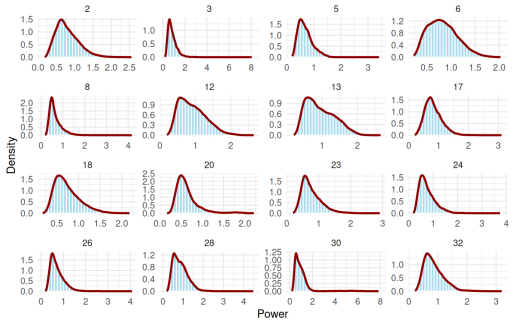
(a) Offset



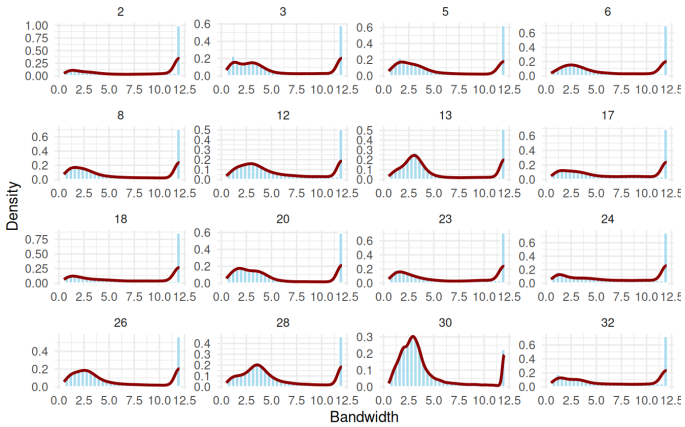
(b) Exponent



(c) Center Frequency



(d) Power



(e) Bandwidth

Figure 3.5: Distribution of FOOOF features across subjects. (a) Offset, (b) Exponent, (c) Center Frequency, (d) Power, and (e) Bandwidth. Histograms with overlaid density curves show the empirical variability across windows.

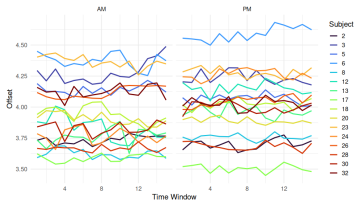
Within-segment Temporal Drift

Inspection of trajectories across the 15 consecutive 20-s windows within each 5-min segment revealed systematic temporal drift in all FOOOF features. At the individual subject level, some trajectories remained relatively flat while others showed clear rises or fluctuations across the segment. When averaging across subjects, a consistent convex rise–then–plateau shape emerged, well captured by a second-degree polynomial in time.

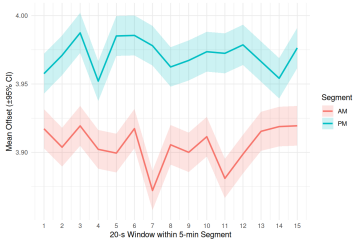
Across all five spectral parameters, clear intra-segment temporal structure was observed (Fig.3.6a–j). For the aperiodic offset (Panels3.6a, 3.6b), individual subjects showed diverse trajectories, ranging from nearly flat to steadily rising or falling profiles. At the group level, both AM and PM segments exhibited a convex drift, with systematically higher offsets in the PM, consistent with a diurnal elevation in broadband log-power. The exponent followed a related but more monotonic pattern (Panels 3.6c, 3.6d), with most subjects maintaining stable within-segment trends; however, group averages revealed consistently higher exponents in the PM, indicating steeper spectral slopes later in the day. Periodic features also showed temporal dynamics. Bandwidth trajectories (Panels 3.6e, 3.6f) varied markedly across individuals but converged to stable means, with AM values slightly exceeding PM and suggesting subtle morning-to-evening narrowing of oscillatory peaks. Center frequency (Panels 3.6g, 3.6h) was more heterogeneous across subjects, yet the group-level means diverged over time: AM frequencies tended to drift downward, whereas PM remained elevated, pointing to a diurnal upward shift. Peak power (Panels 3.6i, 3.6j) was particularly variable, with some subjects exhibiting abrupt within-segment drops. Nevertheless, averages demonstrated systematically higher power in the AM compared to the PM, implying that oscillatory strength was reduced later in the day. Taken together, these results indicate that both aperiodic and periodic spectral parameters undergo consistent short-term drifts within segments and systematic diurnal shifts between AM and PM recordings.

3.3.2 Dependence among FOOOF Features

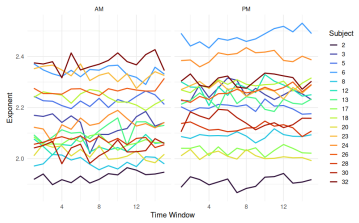
To quantify dependencies between spectral parameters, we computed pairwise correlations across all windows and subjects (Fig. 3.7). The most pronounced association was observed between offset and exponent ($r = 0.72$), consistent with the known coupling between baseline spectral power and the slope of the aperiodic component. This strong dependence indicates that these two features are not fully independent and should be modeled jointly where appropriate. In contrast, offset showed only weak correlations with periodic parameters (center frequency, power, and bandwidth; $|r| < 0.1$). Exponent was nearly uncorrelated with the periodic features ($r \approx 0$), suggesting that slope variability captures a distinct source of variance from oscillatory activity. Among the periodic parameters, center frequency was negatively correlated with power ($r = -0.23$) and positively correlated with bandwidth ($r = 0.16$), pointing to a trade-off between oscillatory strength and frequency location. Bandwidth itself showed only minor associations with the other measures. Overall, these results indicate that while offset and exponent are strongly interdependent, periodic features remain largely orthogonal to the aperiodic components, justifying their separate treatment in the modeling framework.



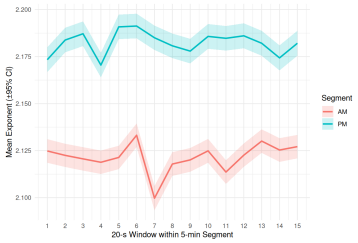
(a) Offset by subjects



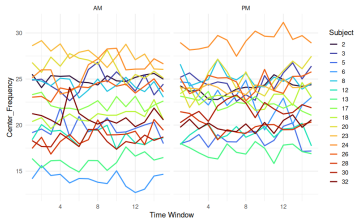
(b) Offset — mean (±95% CI)



(c) Exponent by subjects

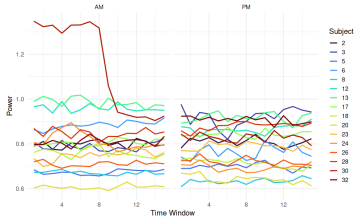


(d) Exponent — mean (±95% CI)

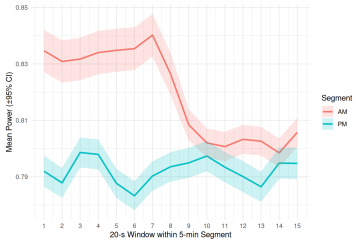


(e) Center frequency by subjects

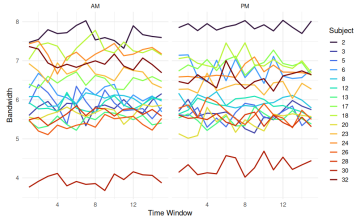
(f) Center frequency — mean (±95% CI)



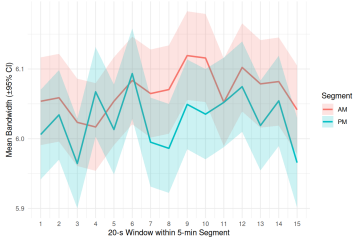
(g) Power by subjects



(h) Power — mean (±95% CI)



(i) Bandwidth by subjects



(j) Bandwidth — mean (±95% CI)

Figure 3.6: Within-segment temporal drift across five FOOOF features. Panels (a,b) Offset, (c,d) Exponent, (e,f) Center frequency, (g,h) Power, and (i,j) Bandwidth. Left column: subject-level trajectories across 15 consecutive 20-s windows for AM and PM segments. Right column: group-level means with 95% confidence bands.

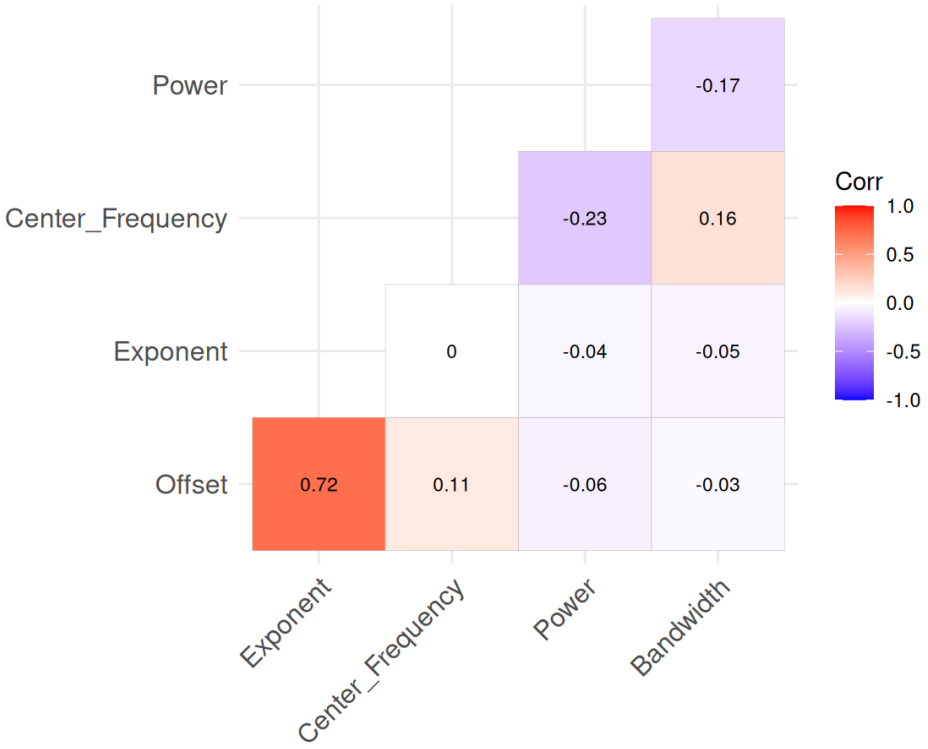


Figure 3.7: Pairwise correlations among FOOOF features. Offset and exponent were strongly positively correlated ($r = 0.72$), while correlations between aperiodic and periodic parameters were weak. Center frequency showed modest trade-offs with power and bandwidth, whereas bandwidth exhibited only minor associations overall.

3.3.3 Stability Mapping via CV

To determine whether brain regions can be stratified by the stability of their spectral features, we calculated the CV for each FOOOF parameter across regions and segments. Figure 3.8 shows the rank ordering of regions by their median CV, separately for Offset, Exponent, Center Frequency, Power, and Bandwidth.

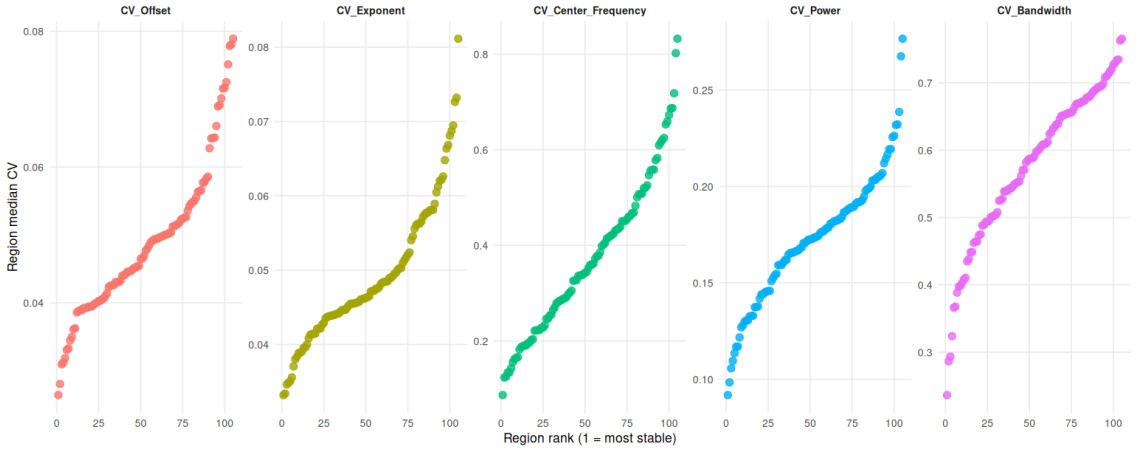


Figure 3.8: Stability mapping of FOOOF features by CV. Each panel corresponds to one spectral feature (Offset, Exponent, Center Frequency, Power, Bandwidth). Regions are ranked by their median CV, with lower values indicating higher stability. The stratification highlights that stability is feature-dependent, with Offset and Exponent being most stable, Center Frequency and Bandwidth showing greater variability, and Power falling in between.

The results reveal systematic feature-specific differences. Offset and Exponent exhibit the lowest CVs across regions, indicating that aperiodic components are the most stable features within and across regions. In contrast, Center Frequency and Bandwidth display substantially higher variability, suggesting that periodic parameters are more sensitive to temporal or regional fluctuations. Power shows intermediate stability, with certain regions clustering among the more stable but others displaying larger dispersion. These results show that stability differs by feature and region, highlighting the need to account for regional variability in subsequent analyses.

Having established that stability, quantified via the CV, varies systematically across features and regions, the next step was to formally test whether these differences are statistically significant across regions. To do so, we employed the Friedman test, a non-parametric alternative to repeated-measures ANOVA that is well suited for ranked data and does not rely on normality assumptions. In addition, we quantified the degree of concordance among regions using Kendall's W, which provides an interpretable effect size for the strength of agreement. Together, these tests allow us to evaluate whether regional stratification reflects systematic heterogeneity rather than random variation.

Table 3.1 summarizes the outcomes of the Friedman test across subjects and features. Across all five spectral features, the majority of subjects showed significant heterogeneity in regional stability. Center Frequency exhibited the strongest and most consistent effect, with every subject (100%) classified as heterogeneous. Offset and Bandwidth followed closely, with heterogeneous decisions in over 93% of subjects. Exponent and Power also

Table 3.1: Summary of Friedman test results for stability (CV) across regions. Each feature is classified as showing either heterogeneous or inconclusive stability patterns. The columns report the number of subjects in each category and the corresponding proportion.

CV of Features	Decision	Subjects	Proportion
Offset	Heterogeneous	15	0.938
	Inconclusive	1	0.062
Exponent	Heterogeneous	14	0.875
	Inconclusive	2	0.125
Center Frequency	Heterogeneous	16	1.000
Power	Heterogeneous	13	0.812
	Inconclusive	3	0.188
Bandwidth	Heterogeneous	15	0.938
	Inconclusive	1	0.062

demonstrated strong evidence of heterogeneity, although a small fraction of subjects were classified as inconclusive. These findings indicate that regional differences in stability are robust and consistently present across individuals.

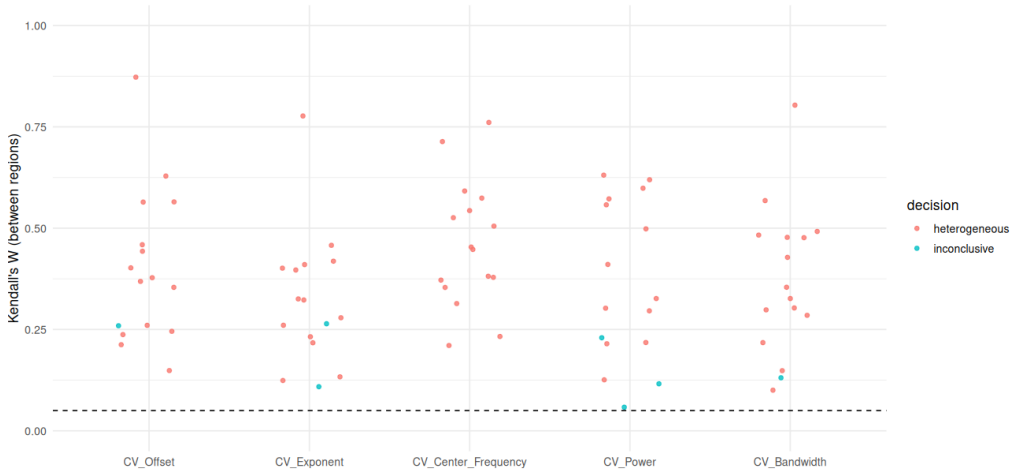


Figure 3.9: Between-region heterogeneity in stability quantified via Kendall's W. Each point represents one subject for a given feature. Colors indicate whether the Friedman test classified the pattern as heterogeneous (red) or inconclusive (blue). The dashed line marks $W = 0$, corresponding to no concordance across regions.

Figure 3.9 provides a subject-level view of regional heterogeneity, expressed through Kendall's W. Values above zero indicate concordance, with larger values reflecting stronger agreement in the rank ordering of regions. The distribution of points reveals that most subjects exhibit substantial heterogeneity across regions for all features. Consistent with the Friedman test, Center Frequency showed the highest and most consistent W values, indicating strong stratification across regions. Offset and Bandwidth also demonstrated robust heterogeneity, whereas Exponent and Power showed somewhat greater variability across subjects, with a few cases falling into the inconclusive range. Together, the statistical test and concordance estimates confirm that regional stratification of stability is both

systematic and replicable at the individual subject level.

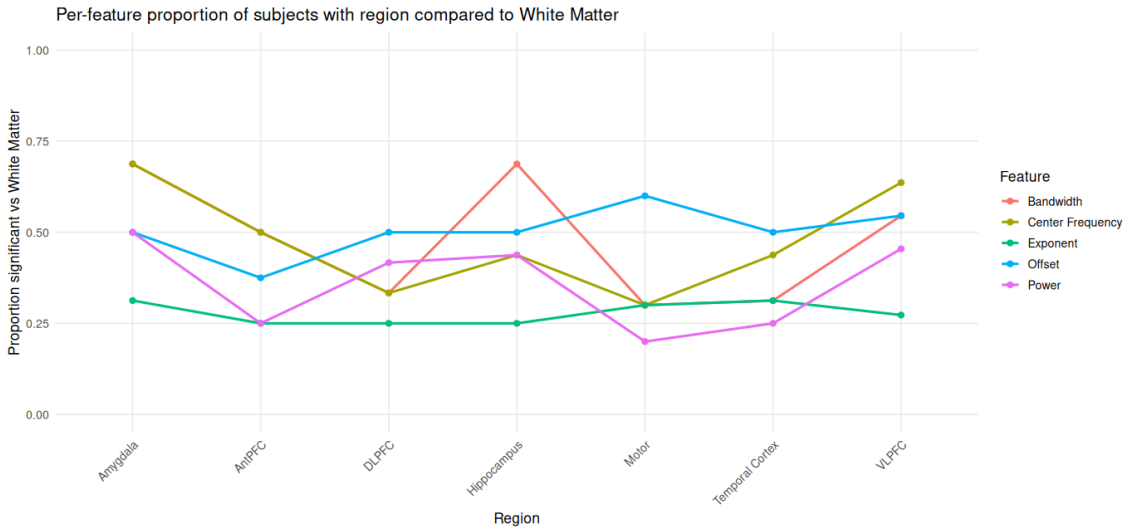


Figure 3.10: Per–feature proportion of subjects with a significant difference between each region and white matter. Each line traces one FOOOF feature (Bandwidth, Center Frequency, Exponent, Offset, Power). Points give the fraction of evaluable subjects for whom the paired Wilcoxon signed–rank test (on shared blocks) was significant at FDR 5% within subject × feature.

Figure 3.10 summarizes the *post–hoc* results by showing, for each region, the proportion of subjects in whom that region's stability (CV) differed from white matter (paired Wilcoxon on shared blocks with BH adjustment). A clear feature hierarchy emerges. *Center Frequency* most often distinguishes gray matter from white matter—particularly in the amygdala and ventrolateral PFC—while *Offset* also shows frequent separation across many regions. By contrast, *Exponent* yields the fewest rejections and *Power* is intermediate. Notably, *Bandwidth* peaks in the hippocampus (and to a lesser extent VLPFC), consistent with broader narrowband peaks in these structures. Motor cortex shows comparatively few differences from white matter except for Offset. These proportions reflect how *often* a feature separates a region from white matter across subjects rather than the magnitude of any single subject's effect, complementing the Friedman layer by indicating where post–hoc contrasts are most reliable.

3.3.4 Hierarchical Model Results

In this stage of the analysis, we examined whether the target regions (amygdala and hippocampus) differed from white matter across all five FOOOF features. The models included fixed effects for recording segment (AM vs. PM) and a quadratic (time^2) term to capture the within-segment drifts observed previously. To account for the repeated structure of the data, we specified random intercepts for subjects and for channels nested within subjects. This hierarchical specification allows us to estimate region-level effects while properly accommodating subject-specific and channel-specific variability.

Univariate Models for Center Frequency and Power

We first report the results of the two univariate models fitted to the periodic features: center frequency and power. Both outcomes required distinct likelihood families: a Gamma distribution with a log link for center frequency, which is strictly positive and skewed, and a Gaussian distribution with a log link for power, where residual variation is well approximated by a normal distribution on the transformed scale. In each case, the model included region, segment, and quadratic time effects as fixed terms, together with random intercepts for subjects and channels. This structure provides baseline assessments of regional differences for the two periodic parameters before turning to the multivariate analysis of Offset and Exponent.

Center Frequency Model

Table 3.2 reports posterior summaries for the univariate Gamma model of center frequency. Estimates are presented as multiplicative ratios (MR), obtained by exponentiating the log-scale coefficients. Values above one indicate an increase relative to the reference category, while values below one indicate a decrease.

Table 3.2: Posterior multiplicative ratios ($MR = \exp\{\beta\}$) for the center frequency model. Values are posterior means with 95% credible intervals.

Parameter	MR	95% Credible Interval (L–U)	$\hat{R}$
Fixed effects			
Segment (PM)	1.02	[1.02, 1.03]	1.00
poly(Time, 1)	5.55	[0.82, 33.5]	1.00
poly(Time, 2)	0.45	[0.07, 3.06]	1.00
Region (Amygdala)	1.55	[1.42, 1.70]	1.05
Region (Hippocampus)	1.43	[1.35, 1.54]	1.12
Segment (PM) × poly(1)	3.74	[0.26, 56.9]	1.00
Segment (PM) × poly(2)	7.03	[0.52, 96.0]	1.00
Random effects	SD	95% Credible Interval (L–U)	$\hat{R}$
subject intercept	0.16	[0.11, 0.25]	1.03
subject:Channel intercept	0.39	[0.37, 0.41]	1.02
Shape parameter	Estimate	95% Credible Interval (L–U)	$\hat{R}$
Gamma shape	2.14	[2.13, 2.15]	1.00

The multiplicative ratio estimates indicate that the center frequency is substantially higher in the amygdala (MR = 1.55, 95% CrI [1.42, 1.70]) and hippocampus (MR = 1.43, 95% CrI [1.35, 1.54]) compared to white matter, with credible intervals excluding the null value of 1. The effect of segment (PM vs. AM) was small but precisely estimated (MR $\approx$ 1.02). Polynomial time terms and their interactions showed wide uncertainty, with intervals spanning 1, suggesting little evidence for systematic within-segment drift. Random-effect variances revealed that variability is more pronounced at the channel level than at the subject level. Collectively, these results demonstrate strong and consistent regional effects on center frequency, particularly for the amygdala and hippocampus relative to white matter, while temporal effects were weak or inconclusive.

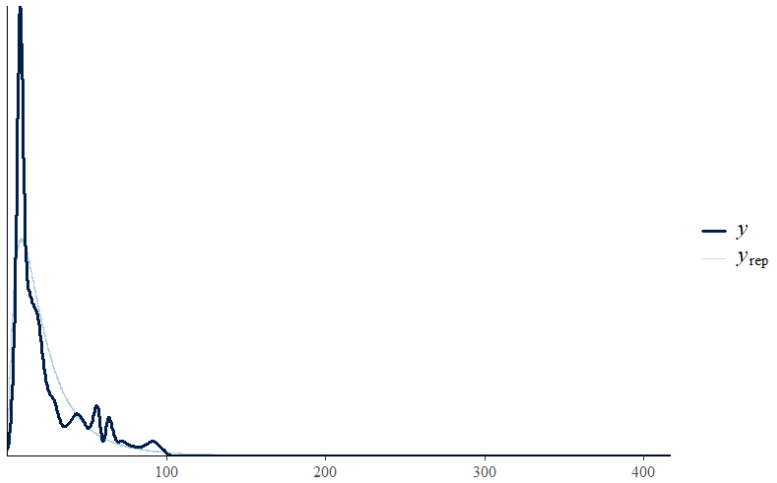


Figure 3.11: Posterior predictive check for the center frequency model. The dark line shows the observed distribution of center frequency values, while the lighter line represents replicated data simulated from the fitted Gamma regression with log link.

To evaluate model fit, we conducted posterior predictive checks by comparing the observed distribution of center frequency values with replicated datasets drawn from the fitted model (Fig. 3.11). The close alignment of both lines indicates that the model adequately captures the skewed distribution of the data, though with a slight underestimation in the extreme upper tail. The model successfully reproduced the strongly right-skewed shape of the empirical distribution, capturing both the sharp concentration of values near lower frequencies and the gradual decline in the upper tail. While the replicated distribution slightly underestimates density at the extreme right, the overall agreement between observed and predicted patterns supports the adequacy of the Gamma specification with a log link for this outcome.

We further assessed model adequacy using approximate leave-one-out cross-validation (LOO) with Pareto-smoothed importance sampling. The diagnostic plot (Fig. 3.12) shows the distribution of Pareto k values across all observations. Nearly all values fall well below the conventional threshold of 0.7, with the bulk clustered around zero. This indicates that the importance weights are stable and that the model generalizes reliably to held-out data. In combination with the posterior predictive checks, these results provide strong evidence that the Gamma specification offers a suitable description of the center frequency distribution.

As an additional check, we computed the WAIC for the center frequency model. The estimated effective number of parameters was $p_{\text{waic}} \approx 1144$, consistent with the model's complexity given the hierarchical random effects. The expected log predictive density (elpd_waic) was stable, and the associated standard errors were modest relative to the magnitude of the estimates. Together with the PSIS-LOO diagnostics, the WAIC results reinforce that the fitted Gamma hierarchical model provides an adequate balance of flexibility and predictive performance, supporting its use for inference on regional differences in center frequency.

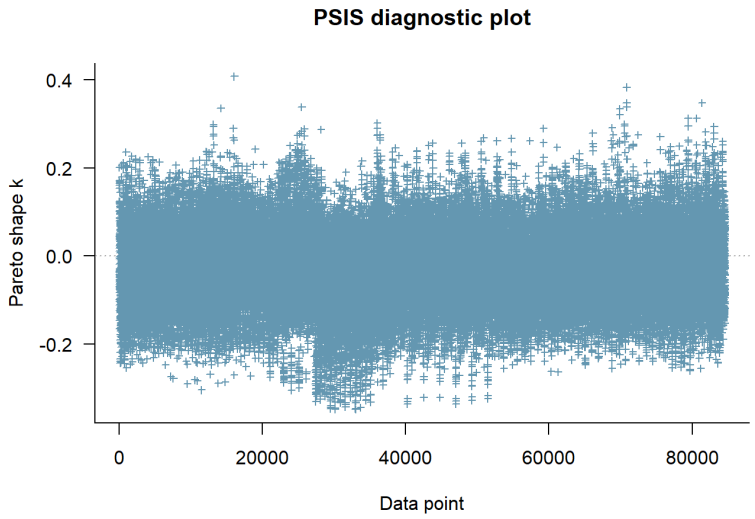


Figure 3.12: PSIS-LOO diagnostic plot for the center frequency model. Each point corresponds to a single observation, with the vertical axis showing the Pareto k estimate of importance weight stability. Nearly all k values fall well below the threshold of 0.7, indicating reliable importance sampling and good out-of-sample predictive performance.

Power Model

Table 3.3 summarizes the univariate lognormal for spectral power (Gaussian residuals on log-power). We report multiplicative effects on the mean ($MR = \exp\{\beta\}$): values >1 indicate an increase relative to the reference (white matter, AM segment), and values <1 indicate a decrease.

Table 3.3: Posterior multiplicative effects ($MR = \exp\{\beta\}$) for the power model. Values are posterior means with 95% credible intervals.

Parameter	MR	95% Credible Interval (L–U)	$\hat{R}$
Fixed effects			
Segment (PM)	0.96	[0.96, 0.97]	1.00
poly(Time, 1)	0.0013	[0.00047, 0.0032]	1.00
poly(Time, 2)	0.30	[0.12, 0.76]	1.00
Region (Amygdala)	0.79	[0.75, 0.84]	1.02
Region (Hippocampus)	0.81	[0.78, 0.84]	1.05
Segment (PM) × poly(1)	1.09×10^{3}	$[2.8\times10^{2}, 4.5\times10^{3}]$	1.00
Segment (PM) × poly(2)	3.60	[0.94, 13.7]	1.00
Random effects	SD	95% Credible Interval (L–U)	$\hat{R}$
subject intercept	0.12	[0.08, 0.19]	1.02
subject:Channel intercept	0.25	[0.24, 0.26]	1.05
Residual SD	Estimate	95% Credible Interval (L–U)	$\hat{R}$
σ (log scale)	0.29	[0.29, 0.29]	1.00

The results indicate lower power in the amygdala (MR = 0.79, 95% CI [0.75, 0.84])

and hippocampus (MR = 0.81, $[0.78, 0.84]$) relative to white matter. PM recordings show a small but precise decrease compared with AM (MR ≈ 0.96). The large negative main effects of the orthogonal polynomial terms, coupled with strong positive PM interactions—especially for the first-order term (MR $\approx 1.1 \times 10^3$; very wide interval)—indicate distinct within-segment temporal trajectories between AM and PM segments rather than a uniform drift. Random-effects estimates show greater variability at the channel-within-subject level than between subjects, and the residual SD on the log scale is moderate ($\sigma \approx 0.29$).

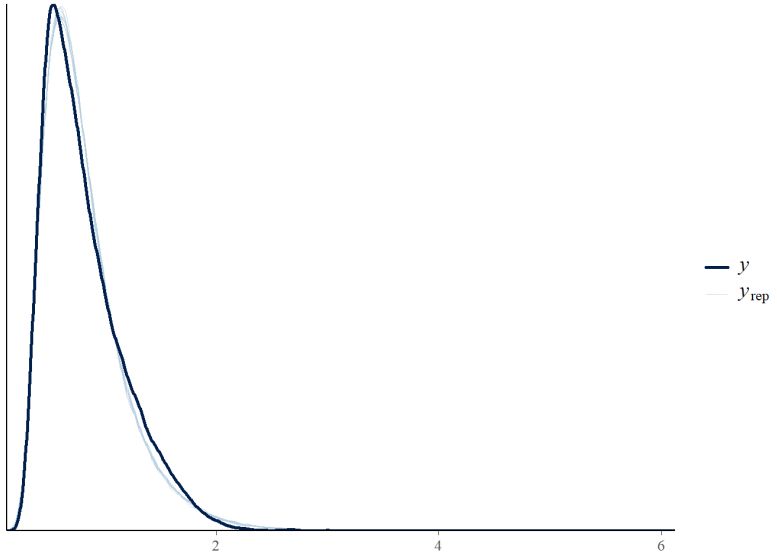


Figure 3.13: Posterior predictive check for the power model. The dark line shows the observed distribution of spectral power, while the lighter line corresponds to replicated datasets simulated from the fitted Gaussian regression with log link.

Posterior predictive checks were also carried out for the power model to assess how well the Gaussian specification with a log link reproduced the observed data distribution (Fig. 3.13). The replicated densities closely followed the empirical distribution, successfully capturing the sharp peak near lower power values and the rapidly decaying right tail. A slight underestimation of the upper tail was visible, but overall, the model reproduced both the central mass and the skewness of the data, indicating that the log-Gaussian formulation provides an adequate fit for spectral power.

We next assessed the predictive adequacy of the power model using Pareto-smoothed importance sampling leave-one-out cross-validation (PSIS-LOO). The diagnostic plot (Fig. 3.14) shows Pareto k values for all observations. All points lie well below the conventional threshold of 0.7, with the majority clustered around zero. This indicates stable importance weights and reliable approximations of leave-one-out predictive densities. Taken together with the posterior predictive checks, the PSIS-LOO results provide strong evidence that the log-Gaussian model generalizes well to held-out data and captures the underlying variability in spectral power.

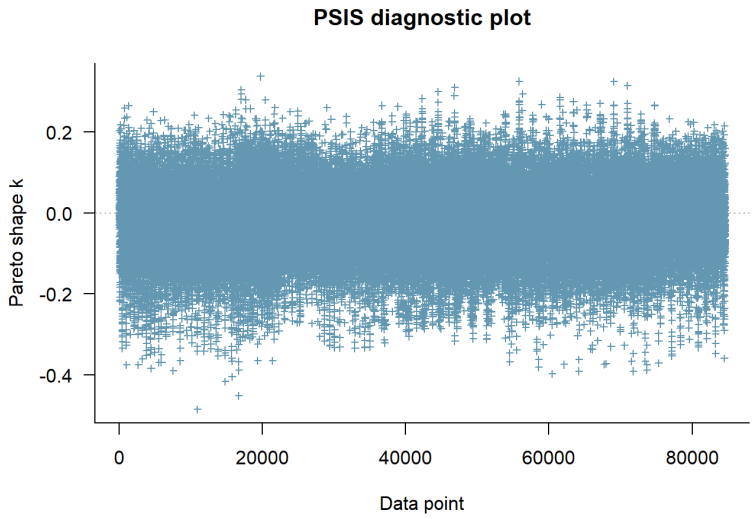


Figure 3.14: PSIS-LOO diagnostic plot for the power model. Each point corresponds to one observation, with the vertical axis showing the Pareto k diagnostic. All k values are well below 0.7, indicating stable importance sampling and good out-of-sample predictive performance.

3.3.5 Multivariate Modeling of Aperiodic Components

We applied a multivariate GLM in a Bayesian hierarchical framework to jointly analyze the aperiodic offset (O) and exponent (B) estimated by FOOOF. The joint specification uses a bivariate Student-t likelihood to accommodate heavy tails while borrowing strength across outcomes, motivated by the strong empirical association between O and B ($r \approx 0.72$). Fixed effects include Segment (AM vs. PM), a quadratic within-segment time trend, and region indicators for amygdala and hippocampus with white matter as reference; Segment-by-time interactions are retained to capture diurnal drift. To respect the data hierarchy, we include random intercepts for subject and for channel nested within subject. Weakly informative priors are placed on regression coefficients, scale parameters, and the residual correlation.

Table 3.4: Fixed effects for the Offset equation on the β (additive) scale with 95% credible intervals. WM = white matter reference; AM is the reference for Segment.

Predictor/Contrast	β	95% Credible Interval (L-U)	$\hat{R}$
Intercept	3.68	[3.53, 3.83]	1.01
Segment (PM)	0.10	[0.09, 0.10]	1.01
poly(Time, 1)	-0.67	[-1.66, 0.30]	1.01
poly(Time, 2)	1.59	[0.58, 2.60]	1.02
Region (Amygdala)	-0.02	[-0.16, 0.10]	1.05
Region (Hippocampus)	0.63	[0.52, 0.74]	1.07
Segment (PM) × poly(Time)[1]	-2.21	[-3.51, -0.80]	1.00
Segment (PM) × poly(Time)[2]	-2.81	[-4.61, -1.43]	1.00

Table 3.4 reports posterior coefficients on the β (additive) scale for the Offset equation. At the center time point, PM segments exhibit a positive effect ($\beta = 0.10$, 95% CrI 0.09–0.10), indicating a reproducible elevation after accounting for subject- and channel-level heterogeneity. Relative to white matter, the hippocampus shows a substantially higher Offset ($\beta = 0.63$, 0.52–0.74), whereas the amygdala is indistinguishable from white matter ($\beta = -0.02$, -0.16–0.10). The within-segment temporal pattern is dominated by curvature: the second-order orthogonal polynomial term is strongly positive ($\beta = 1.59$, 0.58–2.60), while the first-order component is uncertain ($\beta = -0.67$, -1.66–0.30). Notably, PM-by-time interactions are markedly negative ($\beta = -2.21$ and -2.81; 95% CrIs -3.51– -0.80 and -4.61– -1.43, respectively), implying a pronounced attenuation and re-shaping of the linear and quadratic components during PM relative to AM. The intercept is $\beta = 3.68$ (3.53–3.83). These coefficients support a higher baseline Offset in PM and in the hippocampus region, with segment-specific modulation of the within-segment trajectory, all estimated conditional on the hierarchical random effects.

Table 3.5: Fixed effects for the Exponent equation on the β (additive) scale with 95% credible intervals. WM = white matter reference; AM is the reference for Segment.

Predictor/Contrast	β	95% Credible Interval (L -U)	$\hat{R}$
Intercept	2.20	[2.11, 2.29]	1.01
Segment (PM)	0.05	[0.05, 0.06]	1.01
poly(Time, 1)	0.33	[-0.22, 0.89]	1.01
poly(Time, 2)	0.88	[0.33, 1.44]	1.02
Region (Amygdala)	-0.03	[-0.08, 0.01]	1.05
Region (Hippocampus)	0.08	[0.04, 0.11]	1.07
Segment (PM) × poly(Time)[1]	-1.77	[-2.53, -0.99]	1.01
Segment (PM) × poly(Time)[2]	-1.71	[-2.53, -0.92]	1.01

For the Exponent equation (Table 3.5), PM segments show a small but precisely estimated increase at the centered time point ($\beta = 0.05$, 95% CrI 0.05–0.06), indicating a higher aperiodic exponent relative to AM after adjusting for all other terms and random effects. The temporal pattern is again dominated by curvature: the quadratic component is credibly positive ($\beta = 0.88$, 0.33–1.44), whereas the linear component is uncertain ($\beta = 0.33$, -0.22–0.89). PM-by-time interactions are markedly negative ($\beta = -1.77$ and -1.71; 95% CrIs -2.53– -0.99 and -2.53– -0.92, respectively), implying a strong attenuation—and possible re-shaping—of both linear and quadratic trends during PM compared with AM. Regionally, the hippocampus exhibits a modest but credible increase relative to white matter ($\beta = 0.08$, 0.04–0.11), while the amygdala contrast is near-null ($\beta = -0.03$, -0.08–0.01). The intercept is $\beta = 2.20$ (2.11–2.29), representing the baseline exponent for white matter in AM at the centered time point. These coefficients indicate a steeper aperiodic slope in PM and in the hippocampus region, with segment-specific modulation of the within-segment trajectory.

The variance decomposition in Table 3.6 indicates pronounced clustering at the channel level. For the Offset margin, the channel-level intercept SD (0.747; 95% CrI 0.709–0.784) is roughly 2.4 times the residual SD (0.306; 0.305–0.308), corresponding to a $\sim$ 6-fold larger between-channel variance than within-window noise; the subject-level SD (0.258;

Table 3.6: Random-effects and residual standard deviations (SD) with 95% credible intervals.

Grouping	Parameter	SD	2.5% CrI	97.5% CrI
subject	Offset: Intercept	0.258	0.164	0.400
subject	Exponent: Intercept	0.157	0.105	0.234
subject:Channel	Offset: Intercept	0.747	0.709	0.784
subject:Channel	Exponent: Intercept	0.263	0.251	0.278
Residual	Offset (σ)	0.306	0.305	0.308
Residual	Exponent (σ)	0.170	0.169	0.171

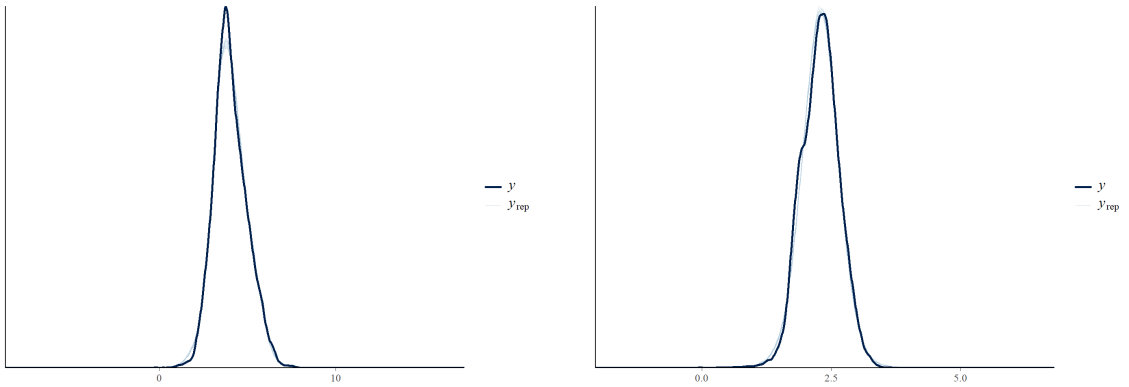
0.164–0.400) is smaller than the residual but non-negligible, confirming heterogeneity across subjects. For the Exponent margin, the channel-level SD (0.263; 0.251–0.278) exceeds the residual SD (0.170; 0.169–0.171) by a factor of ≈ 1.5 in SD units (≈ 2.4 on the variance scale), while the subject-level SD (0.157; 0.105–0.234) is comparable to, though slightly below, the residual. Credible intervals are narrow for residual and channel-level components—reflecting the large number of channel–window observations—and wider at the subject level, consistent with the smaller number of subjects. These patterns justify the nested random-intercept specification (subject; channel within subject) and imply that inference on fixed effects must be conditioned on substantial between-channel variability, particularly for Offset.

Table 3.7: Distributional degrees of freedom and residual correlation between outcomes.

Parameter	Estimate	2.5% CrI	97.5% CrI
Degrees of freedom (ν)	2.23	2.21	2.26
Residual correlation (Offset, Exponent)	0.943	0.942	0.943

The estimated degrees of freedom for the bivariate Student-t likelihood are low ($\nu = 2.23$, 95% CrI 2.21–2.26), indicating exceptionally heavy tails relative to a Gaussian model. This supports the use of a robust error specification, limiting undue influence from outlying windows and accommodating residual heteroscedasticity at the channel level. The residual correlation between Offset and Exponent is extremely high ($\rho = 0.943$, 0.942–0.943) after adjusting for all fixed effects and random intercepts, implying that within-window departures in one outcome are tightly coupled with departures in the other. This strong conditional association justifies joint estimation, improves efficiency for contrasts involving both components, and reinforces the biological interpretation that shared mechanisms underpin concurrent shifts in aperiodic offset and slope.

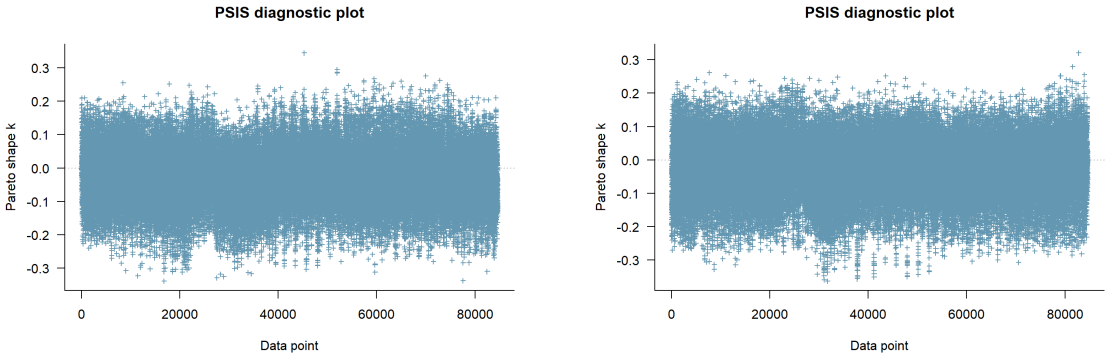
The posterior predictive density overlays in Figure 3.15 show close agreement between the observed and replicated distributions for both margins. For Offset, the model reproduces the modal location and overall dispersion, with replicated draws appearing marginally sharper at the peak—suggesting a slight underestimation of variability near the center but no systematic shift. For Exponent, the central mass is again well captured, with only a modest shortfall of replicated density in the extreme upper tail, consistent with residual heavy–tail behavior. These checks indicate that the bivariate Student-t specifica-



(a) Offset (b) Exponent

Figure 3.15: Posterior predictive density overlays (dark: observed y; light: replicated y_{rep}) for the aperiodic components.

tion provides an adequate marginal fit for both aperiodic components, while any remaining discrepancies are confined to the far tails.



(a) Offset (b) Exponent

Figure 3.16: LOO–posterior predictive overlays comparing observed data (y) with LOO-adjusted replicated draws (y_{rep}) for both margins.

The LOO-adjusted posterior predictive overlays in Figure 3.16 show close alignment between the observed and replicated marginal distributions for both Offset and Exponent. Location and overall spread are well reproduced, with only mild peak sharpening of y_{rep} relative to y and slight under-representation of the far upper tail—more apparent for the Exponent margin. These patterns are consistent with residual heavy-tailed behavior and the estimated low degrees of freedom, but do not indicate systematic miscalibration of central mass or scale. In aggregate, the LOO checks support an adequate marginal fit of the bivariate Student-t specification, with any remaining discrepancy confined to extreme tails.

3.4 Discussion

We combined principled spectral estimation, feature-based reduction, and hierarchical Bayesian modeling to characterize regional differences and temporal dynamics in long-duration iEEG. The pipeline from Welch periodograms with variance reduction, feature extraction with FOOOF, exploratory stability mapping, and multi-level inference yielded a coherent set of results across descriptive and confirmatory stages. In brief, hippocampus showed consistently higher aperiodic level and slope than white matter; periodic features (center frequency and power) displayed region- and segment-dependent patterns with within-segment curvature; and posterior predictive as well as leave-one-out (LOO) diagnostics indicated adequate marginal calibration for all modeled features. Methodologically, aligning each feature with a likelihood that matches its empirical shape (Student-t for Offset/Exponent, Gamma with log link for center frequency, log-normal for power) improved interpretability and predictive validity [7, 106].

Welch's method provided low-variance PSDs per window while maintaining 1 Hz resolution, a practical compromise for 20 s windows; Hamming tapering limited leakage and stabilized peak shapes [79, 81]. Representing each spectrum with FOOOF preserved a clear separation between broadband (Offset, Exponent) and narrowband peaks (center frequency, bandwidth, power), which is advantageous for domain interpretation and for selecting appropriate likelihoods downstream [7]. Empirical histograms reinforced these choices: aperiodic components were roughly symmetric but heavy-tailed, whereas periodic components were positive and right-skewed.

Dependence analyses revealed a strong Offset–Exponent association (motivating joint modeling), modest trade-offs between peak frequency and power/bandwidth, and weak coupling between aperiodic and periodic domains. Stability mapping using the coefficient of variation (CV) showed that Offset and Exponent were the most stable features across regions, with center frequency and bandwidth more labile and power intermediate. Nonparametric, block-aligned comparisons (Friedman tests with Benjamini–Hochberg control, effect sizes via Kendall's W) confirmed heterogeneous regional stability for a subset of subjects, providing an empirical basis for region-focused multilevel models [89, 91, 93].

For the center frequency, a Gamma GLM with a log link reproduced the right-skewed distribution and delivered precise multiplicative effects. Region and Segment effects were credible, with interactions indicating distinct AM/PM trends within-segment time. Power was well captured by a log-normal model; hippocampus and amygdala tended to show lower mean power than white matter, and the PM segments were modestly lower than the AM at the centered time point. Pareto-smoothed importance sampling LOO supported out-of-sample adequacy for both models, and posterior predictive checks replicated modal location and dispersion [106].

Offset and Exponent were modeled jointly using a bivariate Student-t distribution to accommodate heavy tails and to exploit cross-outcome dependence. After adjusting for Segment, quadratic within-segment time, and Region, PM segments exhibited higher Offset and a steeper Exponent at the centered time point. The hippocampus exceeded white matter on both margins, whereas the amygdala contrasts were near-null for the aperiodic parameters. Segment×Time interactions substantially re-shaped within-segment curvature in the PM relative to the AM. The residual correlation remained extremely high ($\rho \approx 0.94$) even after accounting for fixed effects and random intercepts, justifying

the joint specification. Posterior predictive density overlays and LOO-adjusted overlays showed close agreement in location and spread for both margins, with small discrepancies restricted to the far upper tail that are expected under low degrees of freedom [106]. Fitting and inference were carried out in `brms`/Stan, enabling a multilevel structure and robust likelihoods within a unified Bayesian workflow [34, 95].

Random-effect estimates highlighted pronounced channel-level heterogeneity, especially for Offset, which exceeds within-window noise and surpasses subject-level variance. This validates the nested random-intercept structure (subject; channel within subject) and argues against pre-averaging across channels, which could conceal meaningful dispersion. Balanced AM/PM sampling and the second-degree time polynomial capture diurnal level differences and convex drift with minimal complexity, avoiding the aliasing of segment effects into regional contrasts.

We focus on extensions that harmonize with the present results and address the main statistical challenges:

- **Sensitivity to sampling:** (i) Vary the window length $\Delta_{\text{win}} \in \{10, 30\}$ s to study the bias–variance trade-off and its impact on peak estimates and aperiodic parameters; (ii) shift the 5-min AM/PM segments across the 24 h cycle to assess the robustness of diurnal effects; (iii) use a Pareto-LOO grid over the $(\Delta_{\text{win}}, \text{segment})$ lattice to select an optimal blocking scheme based on predictive accuracy [106].

- **Model extensions:** Introduce random slopes (e.g., $1 + t_1 + t_2 \,|\, \text{subject}$) to allow individualized temporal trajectories; extend to multivariate mixed-family models that couple aperiodic and periodic outcomes (Student-t, Gamma, log-normal) via a residual correlation structure or copula, enabling cross-domain contrasts; propagate FOOOF measurement uncertainty with an errors in variables layer when window-level fits are noisy [7].

- **Dynamic hierarchy:** Replace static intercepts with state-space processes $u_p(t)$ and $v_{pc}(t)$ to capture sub-minute dynamics; fit with particle MCMC or sequential Monte Carlo within Stan-compatible interfaces for short-time evolution [34].

- **Spatial regularization:** Smooth channel effects on the electrode graph using ICAR or BYM2 priors to borrow strength across neighboring contacts and stabilize regional contrasts in sparse maps [115].

- **Prior–likelihood conflict checks:** Incorporate conflict diagnostics alongside PSIS summaries to ensure that weakly-informative priors remain compatible as the sample size grows; tailor priors on variance components and residual correlations (e.g., LKJ prior) when partial pooling becomes extreme [104, 106].

- **Computational considerations:** Exploit within-chain parallelization and reduced-rank random effects where feasible; pre-compute frequency-domain quantities and cache FOOOF summaries to shorten iterative refits in cross-validation [95].

A feature-centered, hierarchical Bayesian strategy that is anchored in physiologically meaningful spectral components can scale to hundreds of thousands of window–channel observations while retaining interpretability. In our cohort, the hippocampus shows an elevated aperiodic level and slope relative to white matter; periodic power is reduced in limbic

regions; and diurnal effects modulate both level and curvature within segments. Robust likelihoods, multivariate structure, and LOO-calibrated checks together provide defensible inference and clear directions for refinement. The proposed extensions—sampling-scheme sensitivity analyzes, random-slope and dynamic hierarchies, spatial smoothing, and joint mixed-family modeling—are natural next steps that will deepen physiological insight and strengthen generalization without disrupting the current workflow [7, 95, 106].

Chapter 4

From Spectral Features to Perioperative Dynamics: Change Points Detection and Population-Averaged Modeling of Patient State Index

4.1 Introduction

We adopt a probability–first workflow for perioperative neurophysiology, treating the processed EEG–derived Patient State Index (PSI; 1 Hz) as a time–indexed stochastic process. The analysis targets *instability* rather than only marginal means: (i) detect distributional changes along individual trajectories; (ii) summarize phase–normalized lability with a scale–free Variability Ratio Index (VARI); and (iii) model determinants of instability using population–averaged regression with explicit serial dependence. Inference proceeds via logistic generalized estimating equations (GEE) with a logit link, the binomial variance function $V(\mu) = \mu(1 - \mu)$, and a data–driven working correlation estimated from lag-1 Pearson residual autocorrelations in a two–step procedure, yielding robust, odds–ratio–scale effects while accounting for within–patient dependence [116]. Throughout, we emphasize uncertainty quantification and model criticism with concrete diagnostics.

Signal–level choices reflect the marked nonstationarity of pediatric EEG and the sensitivity of PSI to artifacts and physiologic perturbations. One–second samples are first screened for electromyographic contamination and suppression to reduce artifact–driven variance. Abrupt shifts are then mapped at two levels: at the macro scale, trajectories are segmented with the Pruned Exact Linear Time (PELT) algorithm to locate clinically aligned change points across phases; at the micro scale, Bayesian Structural Time–Series (BSTS) models are fit per child to infer latent level changes with posterior uncertainty [117, 118]. Phase–specific posteriors are converted to VARI, a probability of transition comparable across unequal phase lengths. Covariate associations (e.g., phase, age, sex, BMI, ethnicity where prespecified) with the instant–to–instant probability of a shift are then quantified using GEE with robust (sandwich) errors to guard against working–correlation misspecification [116].

Ensuring adequate—but not excessive—hypnosis is central to pediatric perioperative

safety. Because behavioral endpoints are unreliable in infants and toddlers, clinicians increasingly rely on processed EEG indices, such as the PSI, to titrate anesthetics in real time [119, 120]. PSI aggregates frontal four–channel features (spectral power, phase relationships, interhemispheric coherence) into a 0–100 score, with manufacturer targets of roughly 25–50 during surgical anesthesia [120, 121]. Pediatric EEG, however, is markedly nonstationary: myographic bursts, cardiogenic artifact, hypothermia, and burst–suppression can provoke abrupt excursions [122–124]. Consequently, the raw PSI time series is heterogeneous within and between children, and simple group averages can inflate variance and obscure clinically relevant inflection points [125, 126]. We therefore treat each 1 second PSI observation as a stochastic datum subject to (i) signal–quality fluctuations, (ii) protocol–defined phase shifts, and (iii) latent, patient–specific state transitions.

The analytic pipeline proceeds in four layers. *(1) Quality control.* We exclude epochs with electromyographic activity $> 50\,\mu$V or suppression ratio $> 5\%$, a light–touch filter that removed only 1.6% of seconds but halved within–phase variance, reducing artifact–driven false detections (definitions and clinical context in [122, 123]). *(2) Macro–segmentation.* We locate objective change points using PELT, which computes a global minimizer of a penalized cost,

$$\min_{\{\tau_k\}_{k=1}^{m}} \sum_{k=0}^{m} C\big(y_{(\tau_k+1):\tau_{k+1}}\big) \;+\; \beta\, m, \quad \text{with } \tau_0 = 0,\ \tau_{m+1} = T, \tag{4.1}$$

where $C(\cdot)$ is a segment cost (e.g., negative log–likelihood) and $\beta > 0$ discourages oversegmentation; under standard conditions the dynamic–programming implementation attains linear time in T [127, 128]. *(3) Micro–level modeling.* For each child we fit a BSTS with a local–level state,

$$y_t \;=\; \mu_t + \varepsilon_t, \qquad \mu_t \;=\; \mu_{t-1} + \eta_t, \tag{4.2}$$

optionally with regression and AR components; posterior draws of (μ_t) imply the probability that a genuine state shift occurred at time t [118, 129]. *(4) Phase–normalized instability.* Within each phase p of child i, let S_{ip} be the number of seconds flagged as a credible change point (from BSTS) among N_{ip} seconds. A beta–binomial model treats $S_{ip} \mid \theta_{ip} \sim \text{Binomial}(N_{ip}, \theta_{ip})$ with $\theta_{ip} \sim \text{Beta}(a, b)$ (we use $a = b = 1$ unless otherwise stated), yielding

$$\widehat{\theta}_{ip} \;=\; \mathbb{E}[\theta_{ip} \mid S_{ip}, N_{ip}] \;=\; \frac{S_{ip} + a}{N_{ip} + a + b},$$

a scale–free *Variability Ratio Index (VARI)* comparable across phases of unequal length [130]. Higher VARI denotes a more labile hypnotic state after phase normalization and uncertainty pooling.

To quantify *who* is prone to abrupt PSI shifts and *when* they occur, we model the second–by–second change point indicator $Y_{it} \in \{0, 1\}$ (child i, time t) using GEE,

$$\text{logit}\{\Pr(Y_{it} = 1 \mid \mathbf{X}_{it})\} \;=\; \beta_0 + \mathbf{X}_{it}^{\top}\boldsymbol{\beta},$$

where $\mathbf{X}_{it}$ includes phase, sex, age, BMI, and ethnicity (if prespecified and adequately observed). Serial dependence is handled through a working correlation $\mathbf{R}_i(\alpha)$ and report Huber–White standard errors that remain valid even if $\mathbf{R}_i(\alpha)$ is misspecified [116, 131, 132].

4.2 Materials and Methods

4.2.1 Study Design

We investigate a single-center retrospective study conducted jointly by the Department of Surgical and Vascular Sciences and the Department of Women's and Children's Health at the University of Padua Hospital during the period between January 2021 and October 2021. We included data from pediatric patients' whose ages ranged from 7 days to less than 3 years and who underwent cardiac surgery with extracorporeal circulation (ECC). We also considered only those patients to whom health professionals administer general anesthesia with midazolam and fentanyl and continuously monitor their activity using SEDLINE processed EEG in the operating room. We did not include cases where the SEDLINE sensor is applied to patients with skin lesions, other significant traumatic injuries, or where the medical record information is incomplete. We excluded any datasets with poor-quality processed EEG (pEEG) signals or cases in which patients received propofol during surgery. Table 4.1 presents the baseline characteristics of the selected sample.

Table 4.1: Baseline characteristics of patients

Characteristic	**N**	**N = 20**
Age (in days)	20	176.0 (123.5, 224.0)
Gender	20	
Female		9 (45%)
Male		11 (55%)
BMI	20	14.3 (13.3, 15.4)
Ethnicity	20	
Caucasian		15 (75%)
Asia		2 (10%)
Africa		3 (15%)

Median (Q_1, Q_3); n (%).

PSI data from patients were monitored with SEDLINE and collected in dedicated software (Trace). Retrospectively, patient clinical data preoperatively, during surgery, and postoperatively were collected in an electronic database from paper and digital medical records available at the Hospital. All patients received standard anesthesiological treatment and monitoring based on internal protocols. The stages of surgery were divided, according to clinical practice, into five distinct phases. They are pre-sternotomy phase, open chest pre-pump phase, CPB active hypothermia phase, CPB rewarming phase, and post-pump phase. We retained only one second samples whose EMG artifact was $\leq$ 50% and whose Suppression Ratio was $\leq$ 5% ($\approx$ 98% of the original sample). In total, our dataset comprised 20 pediatric patients and 130,382 one-second observations of PSI. Each pediatric case retains the original, anonymized operating-room identification number assigned in the electronic medical record (e.g., 0, 1, 2, $\cdots$, 23). We deliberately avoided re-labeling patients as '1–20' to preserve the chronological order in which the children entered the study and to facilitate internal cross-checking with source charts. All plots and model outputs, therefore, reference these unchanged ID codes.

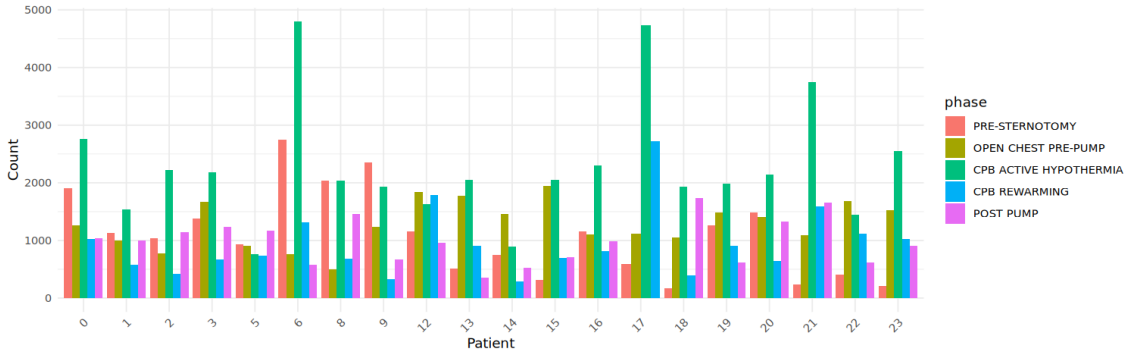


Figure 4.1: Number of Observations per Phase per Patient.

Figure 4.1 shows the distribution of available PSI time points across the five defined anesthesia phases for each patient. Although all patients contributed data to each phase, the number of PSI observations varied considerably between phases and individuals. The CPB active hypothermia phase consistently had the highest density of PSI measurements for most patients, suggesting either longer durations under this condition or more stable acquisition during this intraoperative stage. In contrast, the pre-sternotomy and post-pump phases typically showed fewer PSI time points, which may be due to shorter monitoring durations or preoperative and postoperative interruptions. Some patients, such as patient 6 and patient 17, exhibited a markedly higher number of PSI entries during hypothermia, possibly reflecting individual differences in bypass length or uninterrupted signal acquisition.

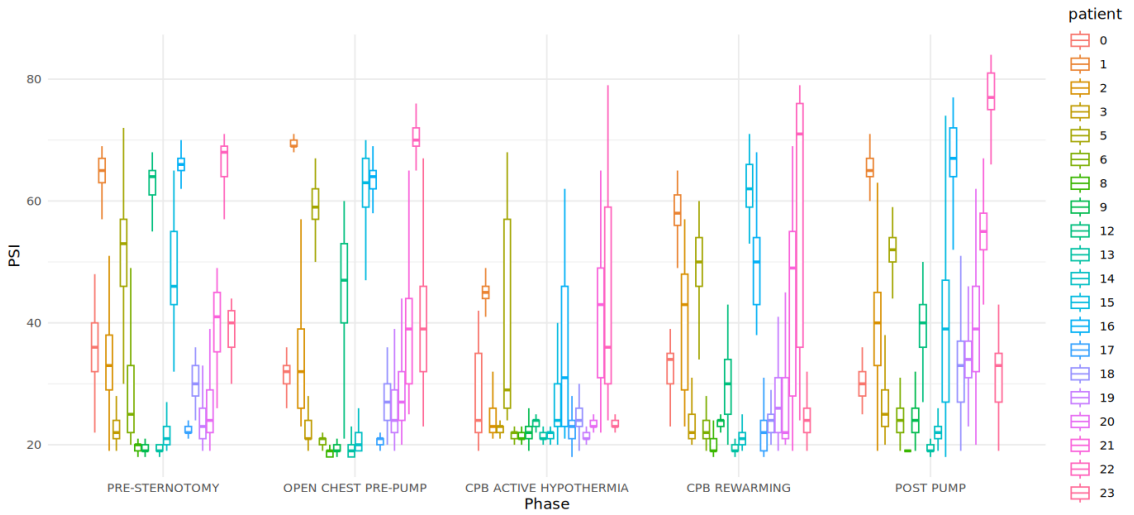


Figure 4.2: PSI across anesthesia phases between patients.

To further characterize intraoperative brain dynamics, we examined the distribution of PSI values across anesthesia phases and individual patients. Figure 4.2 presents patient-wise boxplots of PSI, showing marked variability both across phases and between individuals. While some patients demonstrated a gradual decline and recovery pattern, others showed inconsistent or unexpectedly elevated PSI values during phases typically associated with deep anesthesia or hypothermia. This variability highlights the limited

robustness of PSI when compared across patients and phases, likely due to differences in anesthetic sensitivity, signal quality, and physiological response. These observed differences provided a key motivation for conducting patient-wise analyses in the subsequent stages of this study, where intra-individual EEG dynamics will be explored in greater depth to capture clinically meaningful patterns that may be obscured in aggregate analyses.

Table 4.2: Descriptive summary of PSI across anesthesia phases.

Phase	PSI (mean ± SD)
PRE-STERNOTOMY	34.2 ± 16.9
OPEN CHEST PRE-PUMP	37.9 ± 18.7
CPB ACTIVE HYPOTHERMIA	27.3 ± 10.4
CPB REWARMING	33.3 ± 15.6
POST PUMP	39.6 ± 16.7
p-value (Friedman's ANOVA)	p = < 0.001

Table 4.2 provides a descriptive overview of key EEG-derived parameters (PSI) across the five intraoperative phases. As anticipated, the lowest PSI values were observed during the CPB active hypothermia phase (mean ± SD: 27.3 ± 10.4), reflecting a state of deep anesthetic suppression and hypothermia-induced cortical inactivity. Statistical testing confirmed significant differences across phases (Friedman's ANOVA, $p < 0.001$), underscoring the physiological relevance of these shifts.

4.2.2 Change Point Detection

We study change points in completed perioperative PSI records, so our focus is *offline* (retrospective) detection rather than *online* surveillance. Offline methods search for locations at which the distribution of a series changes; here we target shifts in the mean level of the 1 s PSI process after quality control. Reviews distinguish approximate recursive schemes from exact dynamic–programming approaches [128]. Classical Binary Segmentation (BS) is fast but approximate, while Wild Binary Segmentation (WBS) improves sensitivity to multiple changes [133, 134]. Exact search via dynamic programming delivers solutions that globally minimize a penalized cost but can be expensive; the Pruned Exact Linear Time (PELT) algorithm achieves the same optimum with pruning guarantees and near–linear complexity under standard conditions [127].

Formally, for a PSI sequence $y_{1:T}$ we assume piecewise–constant means and minimize equation 4.1. We use PELT with the modified BIC penalty (MBIC) as implemented in R to control false discovery and enforce a short minimum segment length to avoid single–second spikes [127, 128]. For clinical interpretability, PELT is run independently within each patient and phase, and estimated change point locations are then summarized across patients to identify macro–level shifts tied to surgical stages.

Piecewise–constant models cannot capture gradual drifts. To characterize fine–scale, patient–specific dynamics, we therefore complement PELT with Bayesian Structural Time Series (BSTS) per child using equation 4.2, optionally adding regression and autoregressive terms. Posterior draws of the latent level (μ_t) yield, at each second, the probability that a genuine level transition occurred; these probabilities are propagated rather than

hard–thresholded. Phase–wise instability is then summarized with the *Variability Ratio Index (VARI)*, which pools these second–level probabilities using a beta–binomial model, producing a scale–free measure comparable across phases of unequal duration. This dual design—exact PELT segmentation for macro shifts and BSTS for micro dynamics—keeps inference transparent and clinically aligned while respecting within–patient temporal dependence [118, 129].

Pruned Exact Linear Time (PELT) Algorithm

We use the Pruned Exact Linear Time (PELT) method to segment PSI series into mean–homogeneous pieces. PELT was introduced by Killick, Fearnhead, and Eckley as an *exact* dynamic-programming procedure that returns the global minimizer of a penalized segmentation cost; pruning guarantees yield near-linear complexity under standard conditions [127]. Let $y_{1:T}$ denote the 1 second PSI observations. For a candidate set of change points $0 = \tau_0 < \tau_1 < \cdots < \tau_m < \tau_{m+1} = T$, we minimize

$$\sum_{k=0}^{m} C\big(y_{(\tau_k+1):\tau_{k+1}}\big) + \beta\, m,$$

where $C(\cdot)$ is the segment cost (sum of squared residuals for a Gaussian mean-shift model) and $\beta > 0$ penalizes oversegmentation. PELT evaluates the optimal last-change point recursively, while discarding candidate locations that cannot be optimal because they violate a sufficient inequality,

$$C\big(y_{(u+1):v}\big) + C\big(y_{(v+1):t}\big) + K \;\le\; C\big(y_{(u+1):t}\big),$$

for a constant K determined by the cost; for standard negative log-likelihood costs, $K = 0$ [127]. The resulting pruning does not affect exactness but greatly reduces computation.

For interpretability we run PELT within each patient and phase, using the modified BIC penalty (MBIC) and a short minimum segment length to avoid single-second spikes. Computation is performed in R with the `change point` package [117]. While approximate schemes such as Binary Segmentation and its variants are faster in some settings, PELT provides exact solutions for the chosen cost and penalty and performed reliably on our PSI series [128].

Bayesian Structural Time Series (BSTS) Smoothing

To complement phase-wise PELT segmentation with patient-specific dynamics, we fit a Bayesian Structural Time Series (BSTS) model to each child's PSI trajectory [118, 129]. The observation and state equations are

$$\begin{aligned} y_t &= \mu_t + \mathbf{x}_t^\top \boldsymbol{\beta} + \varepsilon_t, & \varepsilon_t &\sim \mathcal{N}(0, \sigma^2), \\ \mu_t &= \mu_{t-1} + \delta_{t-1} + \eta_t, & \eta_t &\sim \mathcal{N}(0, \tau^2), \\ \delta_t &= \delta_{t-1} + \zeta_t, & \zeta_t &\sim \mathcal{N}(0, \gamma^2), \end{aligned}$$

where μ_t is a local level, δ_t a local trend, and $\mathbf{x}_t$ may include phase indicators or other covariates; $\boldsymbol{\beta}$ are regression effects. The model is estimated via MCMC, yielding posterior draws of the latent level $\{\mu_t\}$ and its first differences $\Delta\mu_t = \mu_t - \mu_{t-1}$.

Rather than thresholding noisy observations, we assess evidence for changes in the smoothed level. At each second we compute

$$p_t \;=\; \Pr(\Delta\mu_t > 0 \mid \text{data})\,,$$

and flag an upward (downward) change point when $p_t > 0.975$ ($p_t < 0.025$). To guard against isolated false positives, we require at least two consecutive flagged seconds. These probabilities are propagated into the phase-level instability summaries (VARI) via the beta–binomial pooling described in the Introduction. Models are fit in R with the `bsts` package [118].

Spectral Motivation for Macro–level Change Points

Ehrenfeld et al. reported distinct frontal EEG spectral patterns during maintenance of general anesthesia, with pattern transitions often aligning with clinically meaningful shifts across intraoperative stages [135]. Although their cohort comprised adults undergoing orthopedic surgery, the observation that spectral content can reorganize at stage boundaries supports our phase-wise search for discontinuities in pediatric PSI. Given children's smaller size and different physiology, we expect higher within-phase variability and more frequent state reconfigurations, which motivates combining phase-level segmentation with patient-level state-space smoothing.

Phase-normalized Instability: the VARI Index

Let $Z_{it} \in \{0, 1\}$ indicate whether a change point is present at second t for patient i within phase p. With $S_{ip} = \sum_t Z_{it}$ change point seconds out of N_{ip} observed seconds, we summarize phase-specific instability by the probability that an arbitrary second in that phase contains a genuine shift,

$$\theta_{ip} = \Pr(Z_{it} = 1 \mid \text{phase } p,\ \text{patient } i).$$

To pool information while remaining scale-free across unequal phase durations, we adopt a beta–binomial model:

$$S_{ip} \mid \theta_{ip} \sim \text{Binomial}(N_{ip}, \theta_{ip}), \qquad \theta_{ip} \sim \text{Beta}(a, b).$$

We set $a = b = 1$ unless otherwise stated (uniform prior). The *Variability Ratio Index (VARI)* for patient i in phase p is the posterior mean

$$\text{VARI}_{ip} = \mathbb{E}[\theta_{ip} \mid S_{ip}, N_{ip}] = \frac{S_{ip} + a}{N_{ip} + a + b},$$

with posterior variance

$$\text{Var}(\theta_{ip} \mid S_{ip}, N_{ip}) = \frac{(S_{ip} + a)\,(N_{ip} - S_{ip} + b)}{(N_{ip} + a + b)^2\,(N_{ip} + a + b + 1)}.$$

This definition keeps VARI in $[0, 1]$, comparable across phases and patients, and automatically accounts for overdispersion beyond simple binomial counting [130]. Uncertainty is propagated via Monte Carlo draws from $\text{Beta}(S_{ip} + a,\ N_{ip} - S_{ip} + b)$ and checked with posterior predictive diagnostics rather than frequentist bootstrap, ensuring consistency with our Bayesian workflow.

Robustness Checking

The robustness of a statistic is a particularly important property, as it refers to the ability to maintain its performance and validity under departures from the assumptions of the statistical model [136]. In other words, a statistic is considered robust if it produces reliable results even when the data deviate from the assumptions made by the statistical method being used [137]. In general, a robust statistic should be insensitive to extreme values or other forms of data contamination and should produce reliable estimates and inferences, even when the data violate some of the assumptions of the statistical method being used. The robustness of a statistic is an important property to consider when analyzing data, especially in situations where the data may contain outliers [138]. By definition, a 'robust' statistic is one that remains reliable even when model assumptions (such as normality) are violated [138]. In this type of research, where the VARI is used, the binary change point variable is expected to have zero inflation over different phases of the data. Therefore, it is important to check the robustness of the VARI statistic.

Resampling methods (e.g., Bootstrap, Jackknife), Monte Carlo simulation, etc., are used to check the robustness of the statistic [138]. The histogram of the statistic is a visual representation of the sampling distribution of the statistic [139]. The sampling distribution is an estimate of the probability distribution of the statistic, which tells how likely it is to observe different values of the statistic if the sampling process is repeated many times [140]. If the histogram of the statistic is normal shape, this suggests that the sampling distribution of the statistic is approximately normal. This is often a desirable property because it allows the use of methods that assume normality, such as confidence intervals and hypothesis tests based on the t distribution [140]. However, it is important to consider that the sampling distribution of the statistic may not always be normal, even if the histogram of the bootstrap samples is bell-shaped. For example, if the original data has a non normal distribution, or if there are outliers or other sources of non normality, the sampling distribution of the statistic may also be non normal. It is common to find the bell-shaped sampling distribution from non-parametric bootstrap resampling techniques [141]. In such a case, it is important to identify the distribution with parameters and use the parametric bootstrap technique to check the robustness. We next assessed the robustness of the VARI metric to ensure our findings are not sensitive to deviations from model assumptions or outliers. Robustness was evaluated using resampling techniques (parametric bootstrap in a Bayesian framework [142] and Monte Carlo simulation [143]) to examine the sampling distribution of VARI. These analyses confirmed that VARI remains consistent (positively skewed and bounded) across 10,000 simulated iterations, suggesting our instability measure is reliable despite the zero-inflation and skew inherent in the change-point data.

4.2.3 Generalized Estimating Equations with Customized Working Correlation

Let $y_{it} \in \{0, 1\}$ denote whether a change point is detected for child/phase cluster $i = 1, \ldots, m$ at second $t = 1, \ldots, n_i$. We model the marginal mean with a logistic link and an explicit first–order Markov term so that the model "remembers" whether a change point

occurred one second ago:

$$\mu_{it} \;=\; \mathbb{E}\big[y_{it} \mid \mathcal{F}_{i,t-1}\big] \;=\; \text{logit}^{-1}(\eta_{it}), \qquad \eta_{it} \;=\; \alpha \;+\; \gamma\, y_{i,t-1} \;+\; \mathbf{x}_{it}^{\top}\boldsymbol{\beta}, \tag{4.3}$$

where $\mathbf{x}_{it}$ collects contemporaneous covariates (lagged change point, age, sex, BMI, ethnicity). Under the Bernoulli variance function $V(\mu) = \mu(1-\mu)$, we allow *phase–level* overdispersion via

$$\text{Var}(y_{it} \mid \mathcal{F}_{i,t-1}) \;=\; \phi_i\, V(\mu_{it}), \qquad \log \phi_i \;=\; \mathbf{z}_i^{\top}\boldsymbol{\theta}, \tag{4.4}$$

with $\mathbf{z}_i$ containing phase indicators (one ϕ_i per patient×phase cluster). Define the diagonal variance matrix $A_i(\boldsymbol{\mu}_i) = \text{diag}\{V(\mu_{i1}), \ldots, V(\mu_{in_i})\}$.

To account for serial dependence within each patient×phase sequence, we use a *cluster–specific, empirically estimated* working correlation R_i and set

$$\Sigma_i \;=\; \phi_i^{1/2}\, A_i^{1/2}\, R_i\, A_i^{1/2}\, \phi_i^{1/2}. \tag{4.5}$$

The customized R_i is built in three transparent steps, using the data themselves:

(i) Pilot fit and Pearson residuals. We first fit the mean model (4.3) under *independence* to obtain fitted means $\hat{\mu}_{it}$ and a dispersion estimate $\hat{\phi}_i$. Pearson residuals are

$$e_{it} \;=\; \frac{y_{it} - \hat{\mu}_{it}}{\sqrt{\hat{\phi}_i\, \hat{\mu}_{it}(1-\hat{\mu}_{it})}}. \tag{4.6}$$

(ii) Within–cluster residual autocorrelations. For each cluster i, we compute empirical autocorrelations at short lags $h = 0, 1, 2, \ldots$:

$$\hat{\rho}_i(h) \;=\; \frac{\displaystyle\sum_{t=h+1}^{n_i} \big(e_{it} - \bar{e}_i\big)\big(e_{i,t-h} - \bar{e}_i\big)}{\displaystyle\sum_{t=1}^{n_i} \big(e_{it} - \bar{e}_i\big)^2}, \qquad \bar{e}_i = \frac{1}{n_i}\sum_{t=1}^{n_i} e_{it}. \tag{4.7}$$

These $\hat{\rho}_i(h)$ summarize the *residual* serial structure that remains after accounting for covariates and the lagged outcome.

(iii) Toeplitz correlation and stabilization. We embed $\{\hat{\rho}_i(h)\}$ into a Toeplitz working correlation,

$$(R_i)_{uv} \;=\; \hat{\rho}_i\big(|u-v|\big), \qquad u, v = 1, \ldots, n_i, \tag{4.8}$$

optionally truncating or tapering beyond a modest lag H (multiply by a smooth weight $w(h) \in [0,1]$) to avoid noisy long–lag entries. If a particular R_i is not numerically positive definite, we apply a minimal ridge

$$R_i^{(\lambda)} \;=\; (1-\lambda)\, R_i \;+\; \lambda\, I_{n_i}, \qquad \lambda \in (0,1) \text{ as small as possible}, \tag{4.9}$$

which preserves the short–lag structure while ensuring stability of Σ_i.

Estimating equations and robust inference: Let $\mathbf{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}^\top$ be the $n_i \times p$ mean–derivative matrix. The estimating equation for $\boldsymbol{\beta}$ is

$$\sum_{i=1}^{m} \mathbf{D}_i^\top \Sigma_i^{-1} \left(\mathbf{y}_i - \boldsymbol{\mu}_i\right) \;=\; \mathbf{0}, \tag{4.10}$$

solved jointly with the dispersion submodel (4.4). Uncertainty is reported using the robust (sandwich) covariance:

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) \;=\; \left[\sum_{i=1}^{m} \mathbf{D}_i^\top \Sigma_i^{-1} \mathbf{D}_i\right]^{-1} \left[\sum_{i=1}^{m} \mathbf{D}_i^\top \Sigma_i^{-1} \left(\mathbf{y}_i - \boldsymbol{\mu}_i\right)\left(\mathbf{y}_i - \boldsymbol{\mu}_i\right)^\top \Sigma_i^{-1} \mathbf{D}_i\right] \left[\sum_{i=1}^{m} \mathbf{D}_i^\top \Sigma_i^{-1} \mathbf{D}_i\right]^{-1}, \tag{4.11}$$

which remains consistent even if the working correlation is mildly misspecified.

At the end, the full pipeline consist of the following steps:

1. **Define clusters and order.** Create clusters as patient×phase sequences, and order observations by time (the `waves` index).

2. **Encode the lag term.** Within each cluster, set $y_{i,0} = 0$ and form $y_{i,t-1}$ so that (4.3) captures short–range dependence directly.

3. **Pilot fit under independence.** Fit (4.3) with independence working correlation to obtain $\hat{\mu}_{it}$ and $\hat{\phi}_i$; compute Pearson residuals e_{it} via (4.6).

4. **Estimate residual ACFs.** For every cluster, compute $\hat{\rho}_i(1), \hat{\rho}_i(2), \ldots$ using (4.7). In practice we retain only short lags (up to the order at which $\hat{\rho}_i(h)$ becomes negligible).

5. **Build R_i (Toeplitz).** Assemble R_i using (4.8); if needed, apply truncation/tapering or the ridge correction (4.9) to ensure positive definiteness.

6. **Fit the final GEE.** Use the fixed, empirically estimated R_i in (4.5) and the dispersion model (4.4) to solve (4.10), and report robust standard errors (4.11).

Clusters are patient×phase (`id = cluster_id`), observations are ordered with `waves = time`, and the lag term $y_{i,t-1}$ is created by within–cluster shifting. The pilot fit uses `geeglm(..., corstr = "independence")` to obtain $\hat{\mu}_{it}$ and Pearson residuals e_{it}. For each cluster we compute the residual ACF (typically retaining short lags), form a Toeplitz R_i, and concatenate its lower–triangular off–diagonal elements into the vector `zcor` required by `geese()`. The final model is estimated. In words: we model the *probability* of a change point at the next second using logistic regression with a lagged indicator, allow overall variability to differ by phase, and supply a realistic within–sequence correlation built from the data; robust (sandwich) errors safeguard inference even if that correlation is only approximate.

4.3 Results

4.3.1 Phase-Wise Change Point Detection

We began by examining changes in sedation levels, measured through PSI, across the five predefined intraoperative phases. By aggregating PSI data across all patients and aligning observations to the surgical timeline, we applied the PELT algorithm to detect shifts in the mean PSI values at phase boundaries.

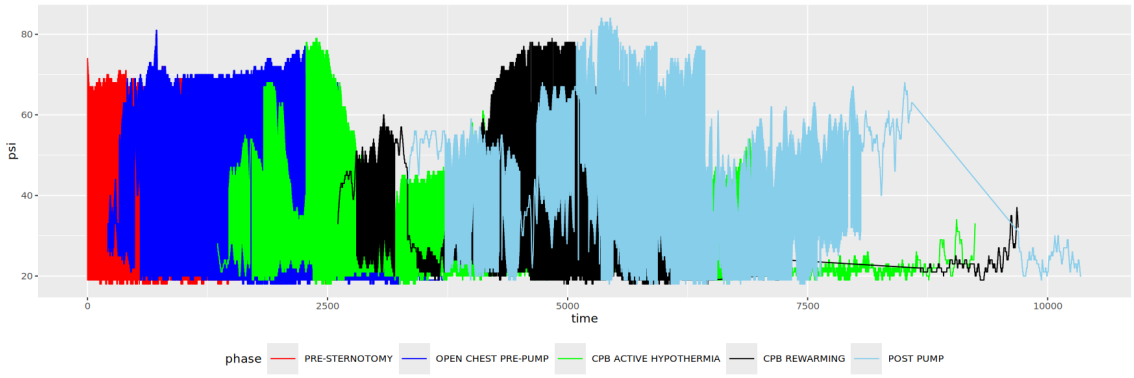
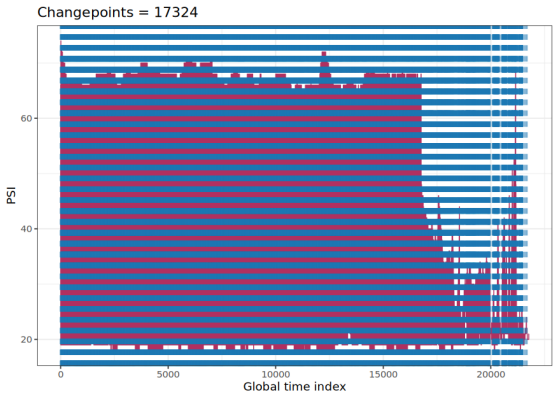


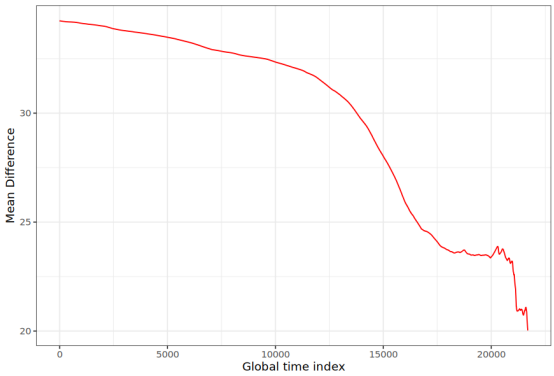
Figure 4.3: Distribution of PSI of 20 patients at different phases over the time.

To see the variation in the PSI value of 20 patients at the end of each phase, we plot the PSI value over the time of their anesthesia period in Figure 4.3. The PSI values are differentiated at each phase by their colors (red for pre-sternotomy, blue for Open Chest Pre Pump, green for CPB active hypothermia, black for CPB rewarming, and sky blue for Post Pump). The sedation value for 20 patients is expected to change at the endpoint of each phase. But the PELT and BSTS algorithms identified multiple change points across the phases of anesthesia. Figure 4.4a - 4.4j displays the change points detected by the PELT algorithm. The plot clearly indicates numerous change points in the PSI of 20 patients within each phase.

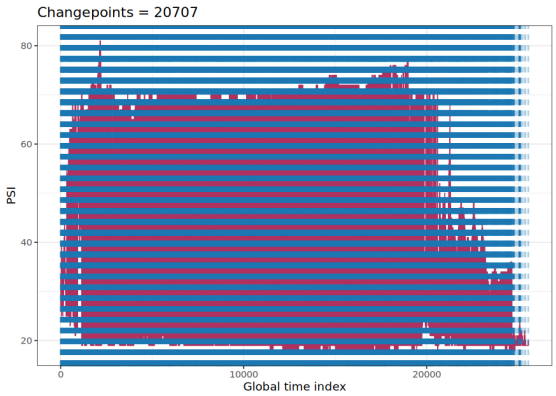
While the pooled-series segmentation is useful for flagging gross macro level transitions, it is intrinsically limited: each concatenation point between children is statistically indistinguishable from a physiological mean shift, so PELT necessarily interprets the handover gap as a change-point. Consequently, the break-density shown in Figure 4.4a - 4.4j conflates true cortical reactivity with artifactual jumps introduced by unequal record lengths, inter-subject baseline differences and sporadic signal loss. To disentangle these sources of variability and to preserve the temporal ordering internal to every child, we therefore re-fit the change point model separately to each cleaned PSI trace using patient-specific Bayesian structural time-series. This micro-level approach eliminates concatenation artefacts, respects autocorrelation, and returns posterior probabilities for genuine neuro-physiological shifts that can be aggregated through the VARI without inflation from record stitching.



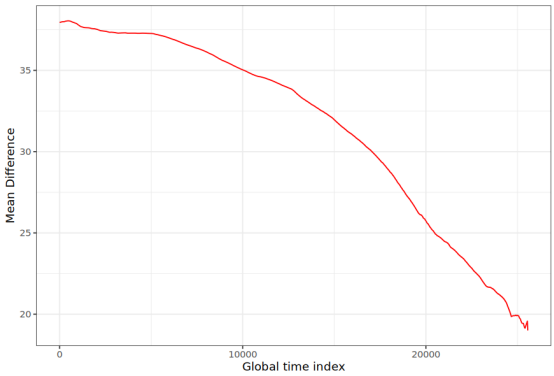
(a) Pre-sternotomy: PSI with detected change points.



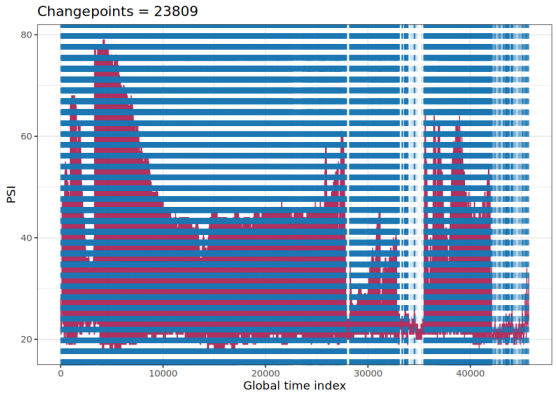
(b) Pre-sternotomy: mean difference at change points.



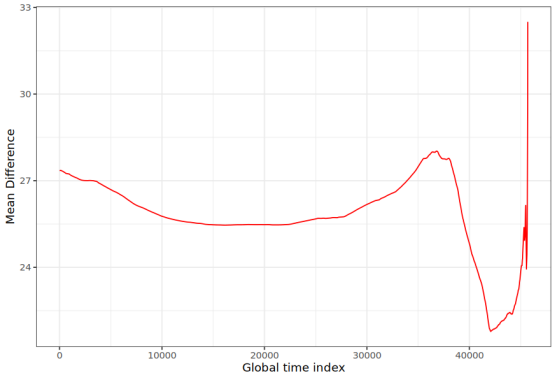
(c) Open chest pre-pump: PSI with detected change points.



(d) Open chest pre-pump: mean difference at change points.

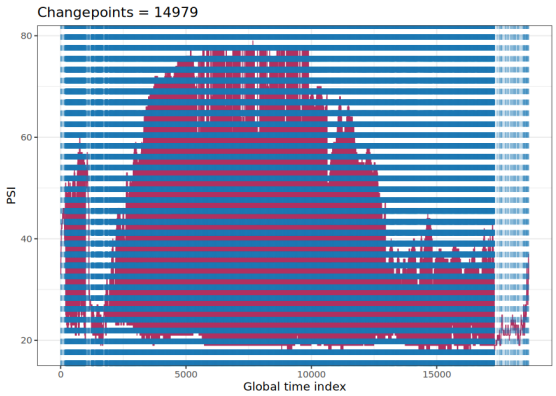


(e) CPB active hypothermia: PSI with detected change points.

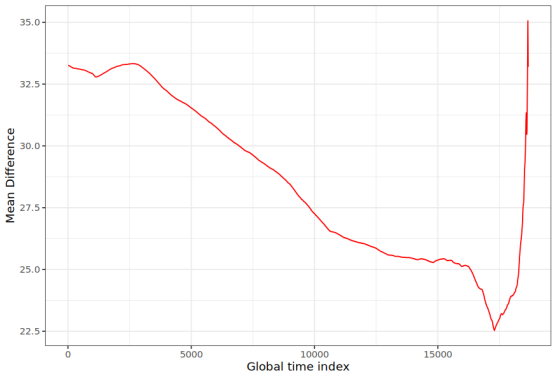


(f) CPB active hypothermia: mean difference at change points.

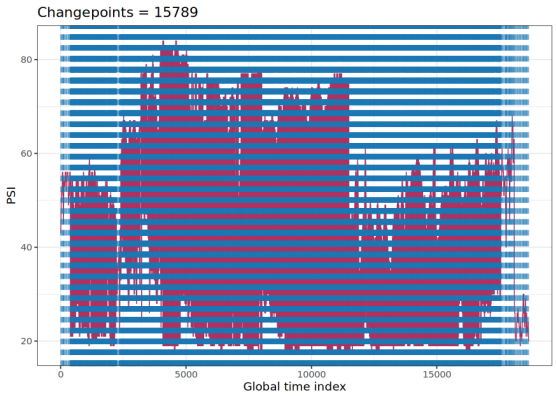
Figure 4.4: (a,b) Pre-sternotomy, (c,d) open chest pre-pump and (e,f) CPB active hypothermia. Left: PSI with vertical sky-blue lines marking change points; right: distribution of the mean PSI difference at those change points.



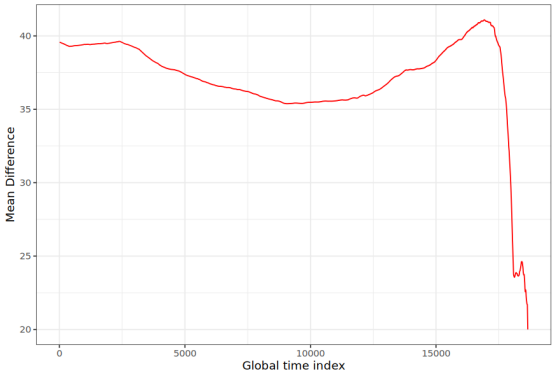
(g) CPB rewarming: PSI with detected change points.



(h) CPB rewarming: mean difference at change points.



(i) Post pump: PSI with detected change points.



(j) Post pump: mean difference at change points.

Figure 4.4: (g,h) CPB rewarming and (i,j) Post pump. Left: PSI with change points; right: mean-difference distribution at change points.

4.3.2 Patient-Wise Change Point Detection

Following the macro-level analysis, we extended change point detection to the individual level using BSTS modeling. For each patient, a univariate state-space model was constructed to represent the PSI time series, incorporating a local level and local linear trend component, along with fixed effects for anesthesia phase indicators. Posterior inference was performed via Markov Chain Monte Carlo sampling over 2000 iterations. The primary focus was on the latent level component, representing the underlying smoothed trajectory of PSI over time. change points were identified by examining first-order differences in the latent level. This approach offers several advantages: it accounts for temporal correlation, accommodates irregular sampling, and provides un-certainty quantification for detected changes. Across patients, the number and timing of change points varied substantially, illus-trating the flexibility of the BSTS model in capturing individualized PSI dynamics that may be overlooked in phase-aligned av-eraging. The variation in change point frequency and location across individuals reflects not only inter-patient heterogeneity but also the potential for detecting latent state transitions that are not strictly aligned with clinical phase demarcations.

Across the 20 BSTS traces in Figure 4.5, the temporal behavior of PSI was strikingly heterogeneous. Roughly one-third of the children (e.g., Patients 1, 8, 14 and 20) exhibited an almost drift-only trajectory with no credible step changes, indicating a stable hypnotic level. On the other side, (Patients 0, 2, 3, 6, 12, 15–19 and 22) the algorithm flagged multiple discrete shifts, typically clustered around marked spikes or troughs in the raw signal. These events were often separated by long stretches of stationarity, suggesting transient cortical arousals or brief artefactual suppressions rather than sustained stage transitions. The remaining subjects (Patients 5, 9, 13, 21 and 23) fell between these extremes, showing one to four change points coincident with isolated excursions of PSI or with slow monotonic drifts that eventually breached the posterior credible limits. Importantly, change points were distributed idiosyncratically along the recording window underscoring the patient-specific nature of cortical responsiveness under deep anesthesia.

This individualized picture contrasts sharply with the phase-wise analysis, where PELT imposed the same five surgical boundaries on every child and therefore ignored within-phase volatility. Aggregating across patients smoothed out most of the short-lived excursions that the BSTS model captured, leading to a systematic underestimation of change frequency in the hypothermic phase and an overestimation in the awakening phase. By conditioning on each subject's own local level, local trend, and observation noise, the BSTS approach retained sensitivity to subtle but clinically relevant deviations, precisely the events that a pooled, phase-locked method is least equipped to detect. Hence, the patient-wise strategy not only yields a more faithful representation of cortical dynamics but also offers a pragmatic route towards personalized anesthetic titration, whereas phase-wise segmentation is better viewed as a coarse descriptive summary for macro-level compliance rather than a decision-support tool.

4.3.3 Comparison of Anesthesia Phases Using VARI

Figure 4.6 displays the distribution of VARI (with 95% confidence interval) at all five phases of this study. As expected, the CPB active hypothermia phase shows the lowest instability (VARI) with the narrowest 95% CI, whereas the post-pump phase shows the

Figure 4.5: Patient-wise PSI change point detection using BSTS (10 patients). Vertical markers indicate detected change points.

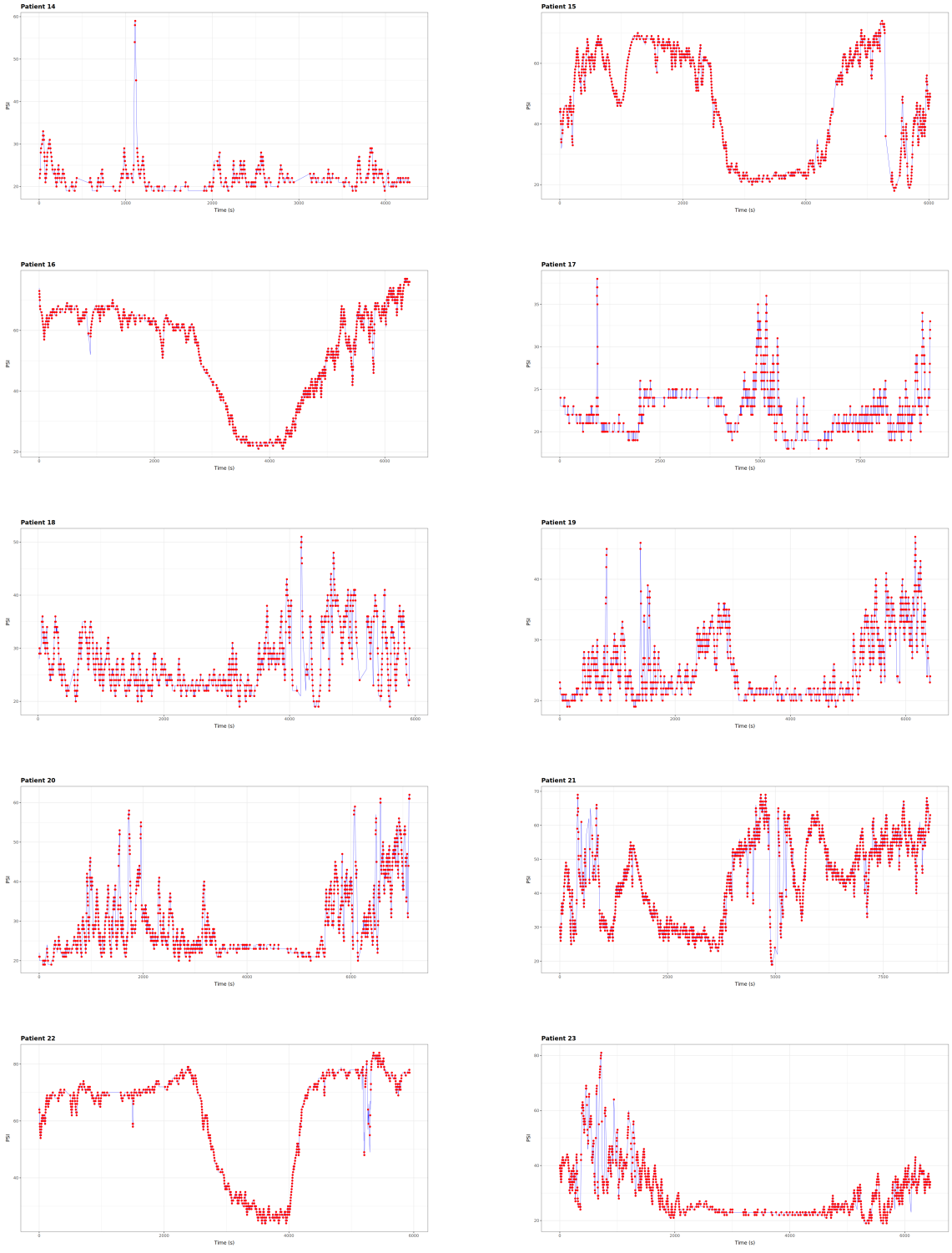


Figure 4.5: Patient-wise PSI change point detection using BSTS (10 patients).

highest instability with the widest 95% CI.

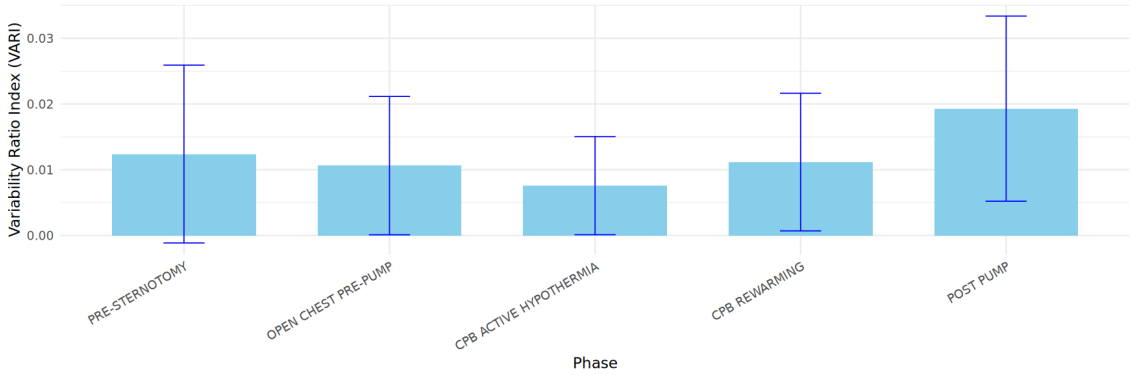


Figure 4.6: Distribution of VARI of mean PSI at separate phases (with BSTS change point detection).

Robustness Check Results

Parametric bootstrapping involves generating multiple datasets from the posterior predictive distribution of the model [142]. This is achieved by drawing parameter values from the posterior distribution and then simulating new datasets based on these parameters, adhering to the assumed parametric form of the data-generating process. Each simulated data set is then analyzed as if it were the original data set, producing a distribution of parameters. To find the distribution of VARI, we first generate the data with the binary 'event' variable, where we insert the value 1 if there is a change in mean PSI at that time point, otherwise 0. So, this variable is generated from binomial distribution with initial probabilities considering as the probability of change points at different phases of the original study data. Then we again calculate the VARI for all phases from the updated data and use these five values as the probability to generate the data for the next iteration. We continue the iteration 10,000 times and plot the histogram with density and box plot of the VARI to identify the distribution. Figure 4.7 displays the histogram with density (left) and the box plot (right) presents the positively skewed shape of the VARI index from the simulated data at all five phases, which preserves the occurrence of zero-change-point observations in the simulated data. The shape of the VARI statistic in all phases converge to a Beta distribution with parameters Shape 1 < Shape 2.

The standard deviation of the simulated VARI is used here to measure the variability. A smaller standard deviation indicates that the simulated VARI statistic is relatively tightly clustered around the mean, whereas a larger standard deviation indicates that the simulated VARI statistic is more spread out. This information helps to understand how much variation in the VARI statistic we might expect to see if this method is applied to other datasets. The graph in Figure 4.8 indicates that the overall variation of the VARI is tightly clustered around the mean, which ensures that the VARI is estimated with high precision across simulations. This would translate into reliable estimates on a new dataset. Specifically, the CPB Active Hypothermia phase (phase 3) shows the lowest variation, as it is the freezing phase.

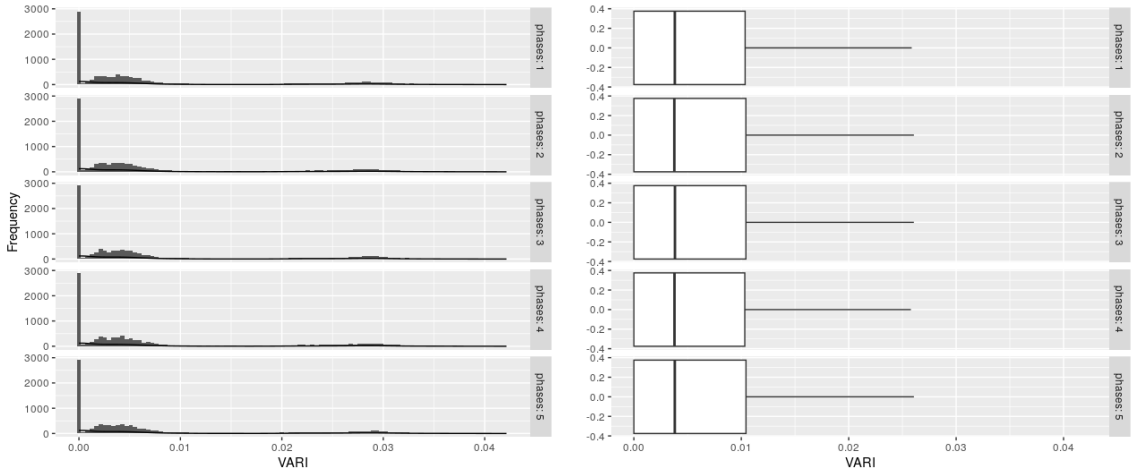


Figure 4.7: Distribution of VARI across different phases: (a) histogram with density of VARI; (b) boxplot of VARI.

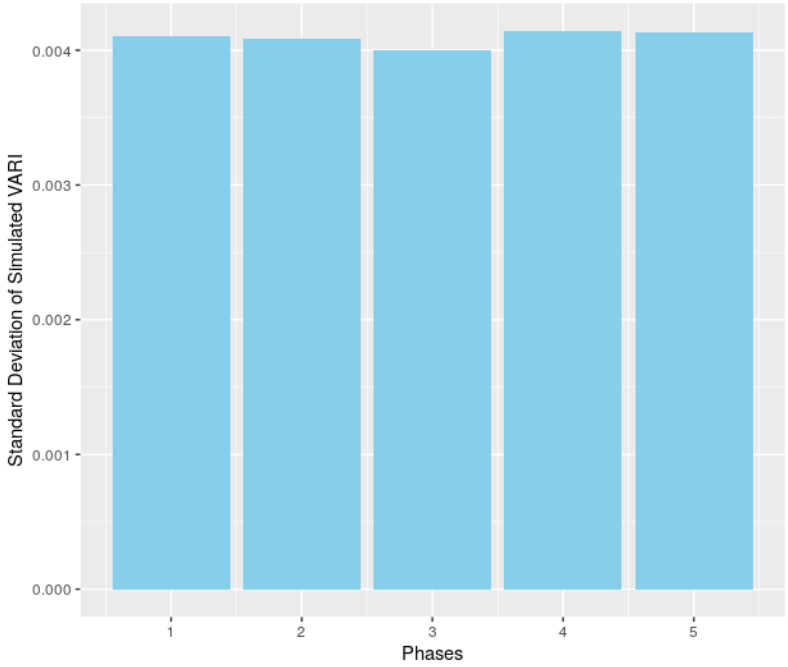


Figure 4.8: Distribution of standard deviation of simulated VARI.

In parametric bootstrapping, we iterate the process 10,000 times. We used beta distribution to generate the VARI. Figure 4.9 demonstrates that the bootstrap estimates (with a 95% confidence interval) of VARI are robust.

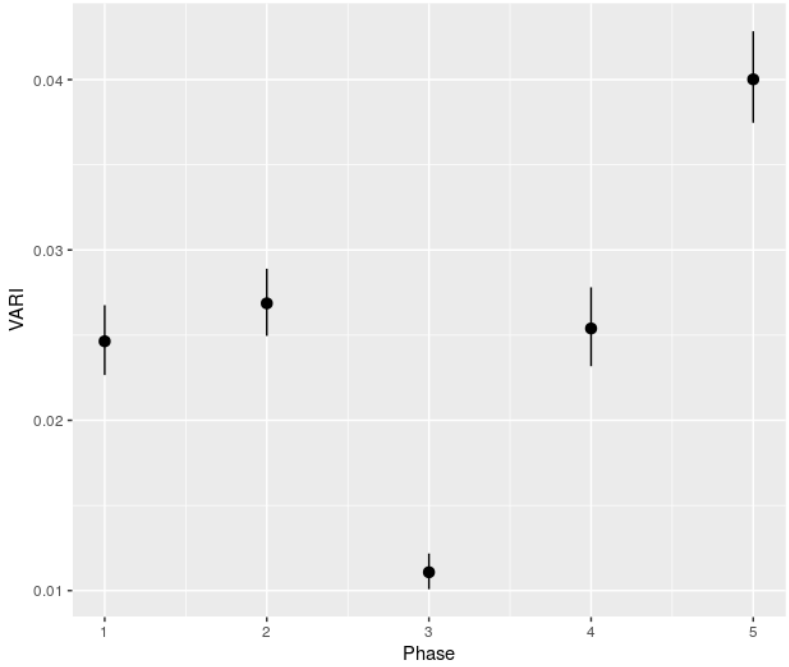


Figure 4.9: VARI (simulated) with 95% confidence interval at distinct phases.

The α and β parameters are used to define the shape of the beta distribution. Specifically, α determines the height of the distribution and β determines the slope [144]. We use a weakly informative prior [145], where the α and β parameters are calculated based on the observed change points and the number of time points in each phase of the data. Figure 4.10 present the bootstrap estimates (10,000 iterations) with the 95% confidence interval for the shape 1 (α on left) and shape 2 (β on right) parameters of the beta distribution of VARI.

In Monte Carlo simulation, we first define the statistic of interest as the VARI in each of the five phases. Then we generated 10,000 simulated data sets that have the same structure as the original data. To do this, we used a zero-inflated beta distribution [146] to generate the data and then calculated the value of VARI statistic. To gain insight into robustness and variability, we used the histogram and the box plot of the simulated statistic here. Figure 4.11 ensures that VARI is relatively robust for all five phases.

4.3.4 Patient-Wise Prediction

Finally, we constructed a predictive model accounting for patient-level factors and within-patient correlations. This population-averaged model incorporates demographic and physiological covariates (age, sex, BMI, ethnicity) and uses a custom correlation structure for repeated measures. By including these patient-specific variables and autocorrelation terms, the model better captures individual variability in PSI responses, although the coefficients represent population-level effects rather than separate per-patient models.

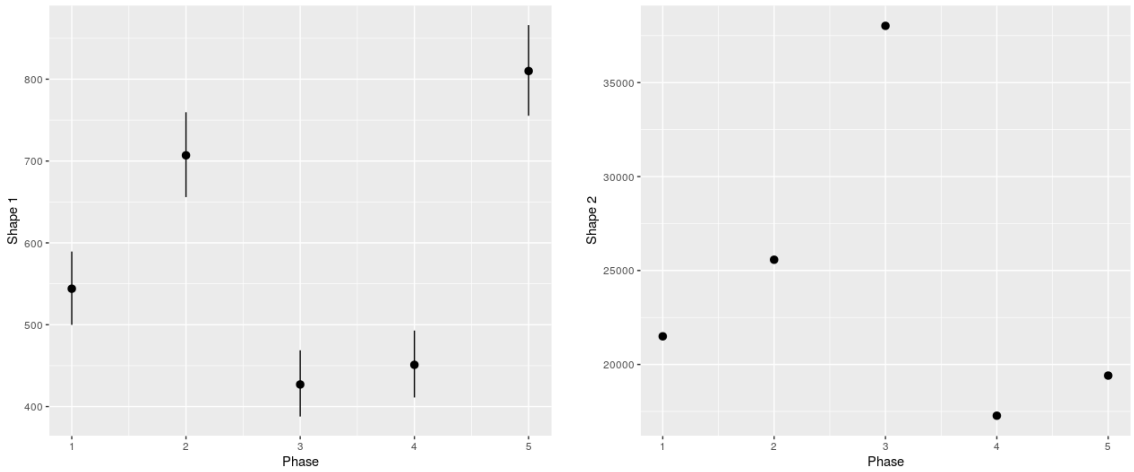


Figure 4.10: Bootstrap estimate with 95% confidence interval for the parameters: (a) shape 1; (b) shape 2.

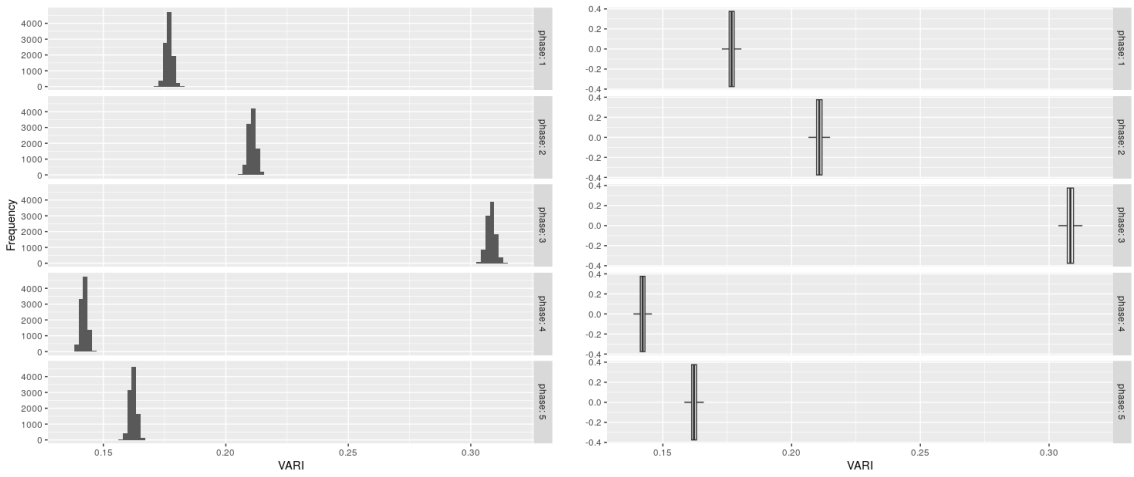


Figure 4.11: Distribution of Monte Carlo-simulated VARIs at different phases: (a) histogram; (b) boxplot.

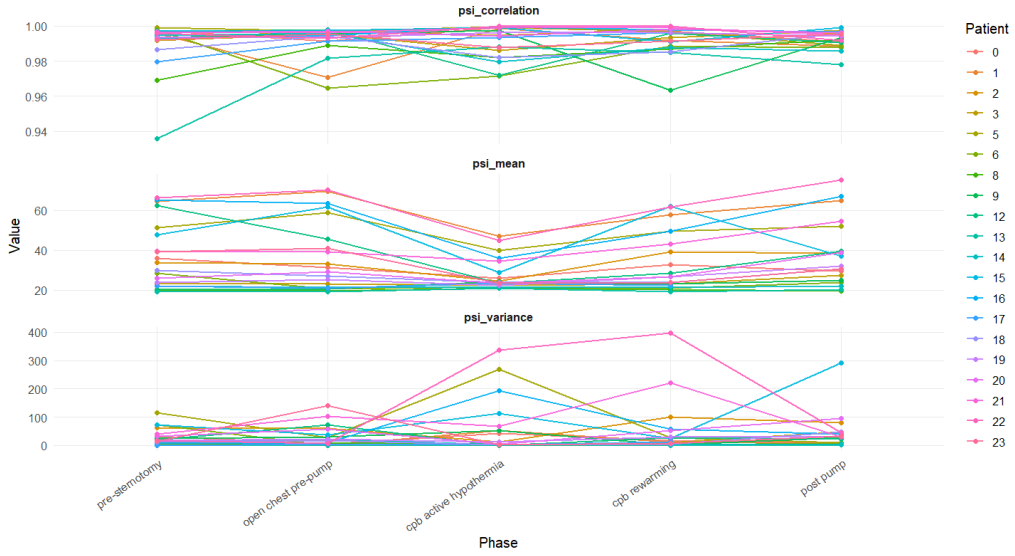


Figure 4.12: Patient wise mean, variance and correlation of PSI at separate phases.

Figure 4.12 illustrates the PSI statistics by patient and phase, providing a comprehensive view of the patient-specific physiological responses. The top panel shows the Pearson correlation between PSI and elapsed time within each phase for each patient, as a simple measure of trend (positive correlations indicate PSI tends to increase during the phase). These correlation values vary by phase, underlining the need for a model that accounts for different temporal patterns. The middle panel shows the mean PSI values for each patient across different phases of anesthesia. It is evident that while some patients maintain relatively stable PSI values, others exhibit significant fluctuations, reflecting the diverse responses to anesthesia. The bottom panel depicts the variance in PSI for each patient and phase, highlighting periods of increased variability. These visualizations underscore the importance of a patient-specific approach in modeling PSI, as they capture the unique and dynamic nature of each patient's physiological state during anesthesia.

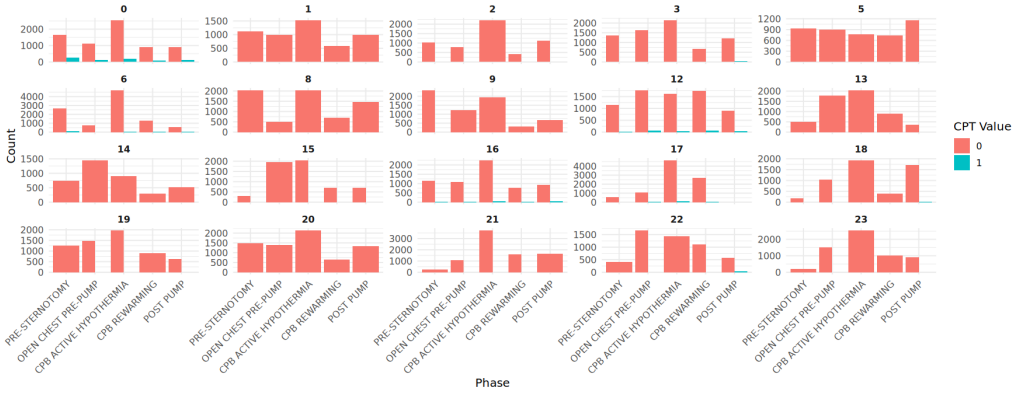


Figure 4.13: change point (CPT) counts by patients and phases.

Figure 4.13 visualizes the distribution of BSTS-derived change points across five anesthesia phases for each of the 20 patients. Each panel corresponds to an individual patient, with bars stratified by phase and colored according to whether a change point

(CPT = 1, cyan) was detected or not (CPT = 0, red). The plots reveal substantial inter-patient variability in both the frequency and timing of PSI shifts. While most time points remain stable (CPT = 0), several patients exhibit elevated change point densities during critical phases such as CPB Active Hypothermia and Rewarming, suggesting increased cortical lability in response to physiological perturbations. This figure supports the clinical intuition that EEG dynamics during surgery are not uniformly distributed but rather modulated by both procedural stages and patient-specific trajectories. Importantly, these change point flags were derived from posterior credible intervals of local-level trends using BSTS, which accommodates autocorrelation and local trends—enhancing sensitivity to subtle but clinically meaningful PSI excursions. By quantifying and modeling these change points with a GEE framework that includes customized correlation, we gain temporally-aware understanding of EEG state transitions during anesthesia.

Table 4.3: GEE with phase as a covariate: odds ratios (OR) for second-wise change point. Robust standard errors, clusters = patient×phase.

Predictor	OR (95% CI)	p-value
Intercept	0.29 (0.13, 0.62)	0.0015
Previous change point (y_{t-1})	0.23 (0.15, 0.35)	< 0.001
Age	0.84 (0.77, 0.93)	0.013
BMI	0.92 (0.87, 0.97)	0.002
Gender: Male (vs Female)	1.06 (0.83, 1.35)	0.640
Ethnicity: Asia (vs Caucasian)	0.85 (0.62, 1.17)	0.318
Ethnicity: Africa (vs Caucasian)	1.08 (0.80, 1.47)	0.607
Phase: Pre-sternotomy (vs CPB active hypothermia)	1.12 (0.84, 1.50)	0.439
Phase: Open chest pre-pump (vs CPB active hypothermia)	1.05 (0.77, 1.45)	0.742
Phase: CPB rewarming (vs CPB active hypothermia)	1.03 (0.80, 1.24)	0.615
Phase: Post pump (vs CPB active hypothermia)	1.29 (1.03, 1.48)	0.052

Baselines: Gender (Female); Ethnicity (Caucasian); Phase (CPB active hypothermia)

Next, we fitted the population-averaged GEE model (described in Section 4.2.3) to quantify how patient factors and phase affect the probability of a change point. The results from the GEE displayed in table 4.3 provide valuable insights into the relationship between various predictors and the probability of change points in the PSI data. In the population–averaged GEE with robust (sandwich) errors and clusters defined as patient × phase, the reference condition is *CPB active hypothermia* with all covariates at their reference levels and no change point at the prior second. The baseline odds of a second-wise change point are low (intercept OR $\approx$ 0.29, 95% CI 0.13–0.62), confirming that abrupt transitions are uncommon once phase and covariates are conditioned upon. A strong history effect is evident: if a change point occurred at $t-1$, the odds of another at t fall by roughly three-quarters (OR $\approx$ 0.23, 0.15–0.35), indicating short-range "refractoriness" in the detection process. Among continuous patient factors, older age and higher BMI are associated with fewer change points after adjustment. Each unit increase corresponds to about a 16% reduction for age (OR $\approx$ 0.84, 0.77–0.93) and a 8% reduction for BMI (OR $\approx$ 0.92, 0.87–0.97). Gender shows no material association (male vs female OR $\approx$ 1.06, 0.83–1.35). Ethnicity terms are imprecise: Asia vs Caucasian is lower (OR $\approx$ 0.85,

0.61–1.17) and Africa vs Caucasian trends higher but remains borderline (OR ≈ 1.08, 0.80–1.47). These patterns suggest that second-wise PSI instability is primarily shaped by immediate neurophysiologic context rather than by gender or broad ethnicity categories, with modest attenuation at older ages and higher BMI within this infant–toddler cohort.

Phase contrasts, taken relative to *CPB active hypothermia*, are clinically coherent. Pre-sternotomy (OR ≈ 1.12, 0.84–1.50) and open-chest pre-pump (OR ≈ 1.05, 0.77–1.45) show borderline-higher odds of change points, whereas rewarming is similar to hypothermia (OR ≈ 1.03, 0.80–1.24). Post-pump shows a trend toward higher instability (OR ≈ 1.29, 95% CI 1.03–1.48, p ≈ 0.05), consistent with physiologic changes during separation from bypass. This suggests a possible uptick in lability after the pump, although the effect is borderline significant.

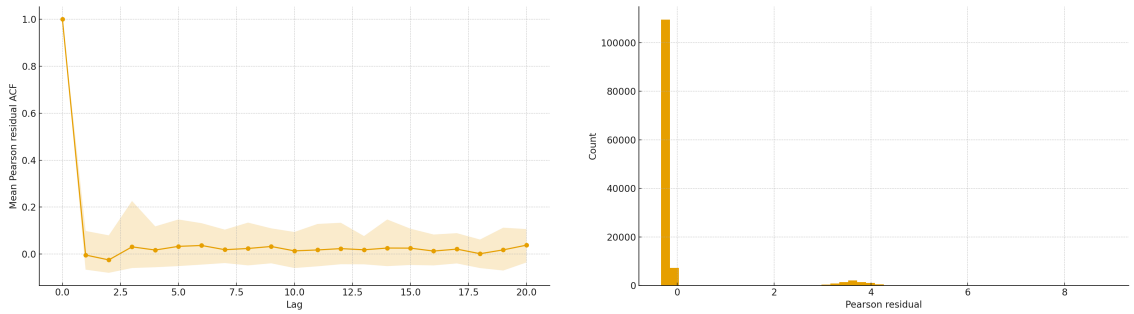


Figure 4.14: GEE diagnostics for the second-wise change point model (clusters = patient×phase; customized data-driven working correlation R_i; robust SEs). **Left:** Histogram of Pearson residuals, concentrated near zero with a small right tail. **Right:** Within-cluster Pearson-residual ACF (mean with 5–95% band) stays near zero beyond lag 0, indicating only modest residual serial correlation.

Figure 4.14 (left) shows the distribution of Pearson residuals from the GEE with customized, data driven working correlation. As expected for a low event–rate binary outcome, the mass is tightly concentrated near zero with a modest right–skew; a small number of positive outliers (residuals > 3) correspond to rare seconds in which change points occurred despite a low predicted probability. This pattern is consistent with a well–calibrated mean model for the bulk of observations and occasional under-prediction at isolated transitions; robust (sandwich) standard errors, as used in Table 4.3, protect inference against the influence of these tails. Figure 4.14 (right) summarizes within–cluster serial correlation of the Pearson residuals: apart from the trivial spike at lag 0, the mean autocorrelation hovers near zero and the 5-95% band straddles zero across lags, with only small positive accents at short lags. This indicates that the explicit lagged change point term captures the dominant short–range dependence and that substantial residual autocorrelation is not evident at the one-second scale. These diagnostics support the adequacy of the mean structure used in Table 4.3 and suggest that the customized, data driven working correlation is not grossly violated for point estimation.

The customized within–cluster correlation R_i from Pearson–residual autocorrelation estimated under an initial independence fit. Across the all patient × phase clusters, residual serial dependence was small: the median (IQR) autocorrelations were $\hat{\rho}(1)$ = -0.021, [-0.038, -0.000], $\hat{\rho}(2)$ = -0.042, [-0.064, -0.021], and $\hat{\rho}(3)$ =-0.012, [-0.036, 0.067]. These values indicate only modest short–lag structure remains after conditioning on the lagged

change point term, so R_i borrows limited correlation while preserving population–averaged interpretability. All cluster–specific Toeplitz matrices were numerically positive definite, so no tapering or ridge shrinkage was applied. Robust (sandwich) standard errors are reported throughout, ensuring valid inference even if the working correlation is mildly misspecified.

4.4 Discussion

This chapter advances a probability–first analysis of perioperative neurophysiology by moving from static spectral summaries to second–wise instability in PSI. Three results anchor the contribution. First, macro–segmentation with PELT provides a coarse, phase–aligned picture of regime changes but is inherently sensitive to record concatenation and inter–subject level shifts. Second, patient–wise BSTS recovers latent, individualized transitions that pooled segmentation systematically blurs. Third, a population–averaged GEE with a customized working correlation links second–wise change points to clinical covariates on an odds–ratio scale, with diagnostics supporting model adequacy. These layers translate raw PSI volatility into phase–normalized probabilities (VARI) and interpretable effect sizes, keeping uncertainty and dependence explicit throughout.

Phase–wise PELT, applied to pooled trajectories, is useful for locating gross changes that coincide with surgical phases; however, it confounds physiologic shifts with stitching artefacts at patient handovers and with unequal series lengths, which inflates break density in some phases and attenuates it in others. This limitation motivates the micro–level analysis. Re–fitting change points per child with BSTS removes concatenation artefacts, respects autocorrelation, and yields posterior probabilities for genuine shifts. In practice, BSTS exposed substantial heterogeneity: several children displayed long stretches of stability, whereas others had clustered, transient excursions coincident with spikes or troughs in the raw signal—precisely the kind of subtle dynamics the pooled, phase–locked view tends to miss. In this sense, macro segmentation is best regarded as a compliance/descriptive tool, while patient–wise state–space modeling offers decision support for individualized titration.

Aggregating BSTS change point posteriors within phase gives the Variability Ratio Index, a scale–free phase probability that is directly comparable across unequal durations. The VARI profile aligns with clinical expectation: instability is lowest (and most tightly estimated) during CPB active hypothermia and highest (with the widest intervals) after separation from bypass. Parametric bootstrapping (10,000 iterations) shows positively skewed sampling distributions across phases that converge to Beta–family shapes; variability is smallest in hypothermia, supporting the interpretation of this phase as "frozen" cortical dynamics. These results indicate that VARI is robust to outliers, skew, and zero–inflation typical of low event–rate binary indicators.

The GEE analysis (clusters = patient × phase) quantifies determinants of second–wise change points while accounting for residual serial dependence via a data–driven, Toeplitz working correlation. Baseline odds of a change point are low (intercept OR $\approx$ 0.29, 95% CI 0.13–0.62), and the strong negative coefficient on the lagged indicator (OR $\approx$ 0.23, 0.15–0.35) suggests short–range "refractoriness": once a change occurs, an immediate repeat at the next second is unlikely. Among continuous covariates, older age and higher BMI are associated with fewer change points (age OR $\approx$ 0.84, 0.77–0.93; BMI OR

$\approx$ 0.92, 0.87–0.97), consistent with damped cortical lability in larger/older infants within this cohort. Phase contrasts are clinically coherent relative to CPB active hypothermia: pre–sternotomy and open–chest pre–pump are similar or only slightly higher, rewarming is comparable, and post–pump is more labile (OR $\approx$ 1.29, 1.03–1.48), matching the physiologic transitions during separation and early recovery. Gender and broad ethnicity categories do not display stable associations.

Model criticism supports these inferences. Pearson residuals concentrate near zero with modest right–skew, and within–cluster residual ACFs hover around zero beyond lag 0, indicating that the explicit lag term captures dominant short–range dependence. The empirically estimated working correlations were small at short lags (median [IQR] $\hat{\rho}(1) \approx -0.021$ [-0.038, -0.000], $\hat{\rho}(2) \approx -0.042$ [-0.064, -0.021], $\hat{\rho}(3) \approx -0.012$ [-0.036, 0.067]); all cluster–specific Toeplitz matrices were positive definite without the need for tapering or ridge shrinkage. Robust (sandwich) errors therefore provide protection even if residual dependence is mildly misspecified.

At the bedside, these results encourage combining a phase–aware macro view (for situational awareness) with micro–level, patient–specific monitoring (for titration). Low baseline odds and the refractory pattern imply that isolated detections should not trigger abrupt interventions without corroboration, whereas sustained elevations in VARI—particularly post–pump—warrant closer review of temperature, perfusion, and hypnotic dosing. The lack of a strong population–wide PSI–instability link after adjustment cautions against interpreting absolute PSI values as sufficient surrogates for lability; dynamics and context matter.

The study is single–center with $n = 20$ children under a specific anesthetic regimen, which constrains generalizability. PSI is a processed index and can be artifact–sensitive despite screening; residual misclassification of change points is possible. BSTS detection thresholds (credible–interval criteria) and the PELT penalty influence event counts; sensitivity analyses to these tuning choices are natural next steps. Finally, while GEE targets population averages, complementary subject–specific models (e.g., Bayesian mixed–effects state–space, joint mean–dispersion formulations) could refine individualized risk predictions. Prospective validation, multicenter replication, and integration with raw EEG features (e.g., aperiodic/periodic markers from Chapter 3) would strengthen external validity and enable real–time decision support. In sum, coupling BSTS–based individualized change points with a phase–normalized instability index and a rigorously diagnosed GEE provides a coherent statistical narrative from physiology to decision scales: hypothermia is stably quiescent, post–pump is labile, and second–wise transitions are shaped more by immediate neurophysiologic context than by crude demographics.

Chapter 5

Discussion

This dissertation advances an integrated analytics paradigm for *Monitoring Patient Reported Outcomes and Neurophysiologic Signals by Integrating Clinical Records with High-Dimensional Digital Data*. Across heterogeneous sources—PRO questionnaires, electronic health records (EHR), and continuous neurophysiologic signals—the central problem is to extract interpretable, decision-relevant summaries while preserving uncertainty and acknowledging dependence structures. Throughout, Frequentist and Bayesian approaches are treated as complementary, not competing, inferential modes: the former offers transparent estimators, hypothesis tests, and well-understood operating characteristics; the latter supplies hierarchical pooling, prior incorporation, and predictive distributions suited to sparse strata and multilevel data [105, 147, 148]. Scalable cohort infrastructures such as the *All of Us* program demonstrate how survey, laboratory, and EHR streams can be harmonized to support population-level inference and precision surveillance [5]. In parallel, modern signal processing now enables the parameterization of high-dimensional EEG into physiologically meaningful features that can be modeled jointly with clinical covariates [7]. The thesis contributions align these currents: Chapter Two develops an integrative toolkit for infectious-disease analysis using GLMs, penalization, hierarchical Bayes, and time-series models; Chapter Three adapts feature-based, hierarchical Bayesian modeling to long-term intracranial EEG (iEEG); and Chapter Four links change-point detection in an EEG-derived anesthetic depth index (PSI) with population-averaged inference via GEE. The result is a coherent route from raw, multi-modal inputs to calibrated, clinically interpretable evidence.

Chapter Two establishes the methodological foundation for surveillance and clinical epidemiology by juxtaposing GLMs and their correlated-data extensions (GLMMs, GEEs) with Bayesian hierarchical models and predictive learners. Penalized regression (e.g., LASSO) and Bayesian shrinkage priors address high-dimensionality and collinearity while preserving interpretability of effect sizes [4, 105]. Model adequacy and generalization are examined with information criteria and cross-validation in the Bayesian workflow [106], while missing-data mechanisms (MCAR/MAR/MNAR) are handled through multiple imputation and fully Bayesian joint models that propagate imputation uncertainty [149, 150]. For policy and stewardship evaluation, the chapter applies classical ARIMA and Bayesian structural approaches for interrupted time series, enabling estimation of both level and slope changes attributable to interventions [151, 152]. The *All of Us* case study illustrates how PRO-like survey responses, EHR covariates, and utilization histories

can be combined to quantify drivers of antibiotic prescribing and to evaluate a national stewardship campaign [5]. In this context, the dual Frequentist–Bayesian lens is pragmatic: likelihood-based estimators provide baselines and diagnostics, whereas hierarchical priors stabilize subgroup estimates and generate posterior probabilities that map directly to decision questions [105, 147, 148].

Chapter Three pivots to high-dimensional neurophysiology and demonstrates that interpretable feature extraction is a gateway to principled modeling. Using FOOOF, each power spectrum is decomposed into an aperiodic component and oscillatory peaks, yielding parameters such as the aperiodic exponent and band-limited power that are linked to plausible physiology [7]. This mimics the role of well-chosen covariates in clinical regression: it reduces dimensionality without sacrificing interpretability. Empirically, stability mapping showed a clear feature hierarchy: aperiodic Offset and Exponent were the most stable, Center Frequency and Bandwidth were more variable, and Power fell in between. Friedman tests confirmed robust between-region heterogeneity in stability across patients; for example, Center Frequency was heterogeneous in 100% of subjects, while Offset and Bandwidth were heterogeneous in 93.8%, and Exponent and Power in 87.5% and 81.2%, respectively. Post-hoc paired comparisons against white matter highlighted that Center Frequency most often distinguished gray-matter regions, with Offset also frequently separating regions; Exponent yielded the fewest rejections. Turning to parametric models, the univariate Gamma model for Center Frequency indicated substantially higher values in the amygdala (MR = 1.55, 95% CrI $[1.42, 1.70]$) and hippocampus (MR = 1.43, $[1.35, 1.54]$) versus white matter, with only a small segment effect (PM vs. AM MR ≈ 1.02) and little evidence for within-segment drift. For Power, a lognormal specification estimated a lower mean power in the amygdala (MR = 0.79, $[0.75, 0.84]$) and hippocampus (MR = 0.81, $[0.78, 0.84]$) than in white matter, a small decrease in PM vs. AM (MR ≈ 0.96), and large AM/PM differences in temporal trajectories captured by orthogonal-polynomial interactions; variability was larger at the channel-within-patient level than between patients, and the residual SD on the log scale was ≈ 0.29. These results suggest further investigation of the power model with interaction effects. Finally, a joint Bayesian multilevel model for the aperiodic Offset–Exponent pair (bivariate Student-t) showed a higher baseline Offset in PM segments ($\beta = 0.10$, 95% CrI $[0.09, 0.10]$) and in the hippocampus vs. white matter ($\beta = 0.63$, $[0.52, 0.74]$), with the amygdala near the white matter baseline; PM×time interactions were strongly negative, reshaping within-segment trajectories. The model captured heavy tails (degrees of freedom $\nu \approx 2.23$) and an extremely high residual correlation between Offset and Exponent ($\rho \approx 0.943$), justifying joint estimation; variance decomposition showed dominant channel-level clustering (e.g., channel-intercept SD ≈ 0.747 vs residual SD ≈ 0.306 for *Offset*). Model checks supported adequacy: posterior predictive overlays matched observed marginals, and PSIS-LOO Pareto-k values were well below 0.7, consistent with a well-calibrated hierarchical fit [105, 106]. The contribution is methodological as well as substantive: parsimonious features plus hierarchical Bayes convert dense iEEG into clinically relevant, uncertainty-aware summaries that enable region-wise comparisons at the population level [7, 105].

Chapter Four addresses perioperative monitoring of anesthetic depth in pediatric patients using the EEG-derived Patient State Index (PSI). Averages can conceal abrupt neurophysiologic transitions; therefore, the analysis treats the PSI trajectory as a time series with latent change points that align with surgical phases and events. When PSI was

pooled across children and segmented phase-by-phase with PELT, the algorithm flagged $\sim$ 14,979–23,809 mean-level shifts per phase, motivating patient-wise modeling. We therefore combined algorithmic detection with Bayesian state-space reasoning to localize individualized shifts and summarize phase-normalized instability via the VARI index; instability was lowest and most tightly estimated during CPB active hypothermia and highest (with the widest intervals) after separation from bypass. Population-level inference then proceeded via generalized estimating equations with a customized, data-driven working correlation [116, 153], yielding low baseline odds of a second-wise change point (intercept OR $\approx$ 0.29, 95% CI 0.13–0.62) and a strong negative lag effect (OR $\approx$ 0.23, 0.15–0.35), consistent with short-range refractoriness. Older age and higher BMI were associated with fewer change points (age OR $\approx$ 0.84, 0.77–0.93; BMI OR $\approx$ 0.92, 0.87–0.97), while gender and broad ethnicity categories showed no stable associations. Phase contrasts (vs CPB active hypothermia) were clinically coherent: pre-sternotomy and open-chest pre-pump were similar or slightly higher, rewarming was comparable, and post-pump was more labile (OR $\approx$ 1.29, 1.03–1.48, $p \approx 0.05$). Diagnostics supported adequacy: Pearson residuals concentrated near zero with a modest right tail, the residual ACF hovered near zero beyond lag 0, and the empirically estimated short-lag correlations were small (e.g., median $\hat{r}(1) \approx -0.021$), indicating that the lag term captured the dominant short-range dependence. This design illustrates how fine-grained signal events (change points) can be scaled to a cohort setting without abandoning marginal (population-averaged) interpretability and shows a clean interface between time-series segmentation (e.g., PELT) and semiparametric correlated-data inference [116, 154]. Clinically, it re-frames perioperative EEG from static thresholds to a probability-first view of state lability that is sensitive to phase transitions.

Taken together, Chapters Two–Four pursue a common strategy: (i) distill heterogeneous inputs into interpretable features (survey responses and curated EHR covariates; aperiodic/oscillatory EEG components; change-point events), (ii) choose likelihoods and estimators that respect data-generating structure (GLM/GLMM/GEE for independence/-correlation patterns; state-space and ARIMA/BSTS for temporal dependence), and (iii) quantify uncertainty via complementary Frequentist and Bayesian devices that support transparent communication and calibrated prediction [105, 106, 147, 148, 152]. Across domains, variability is modeled explicitly rather than treated as noise: heterogeneity in prescribing and intervention response (Chapter Two), regional stability of neural spectra (Chapter Three), and second-to-second volatility in anesthetic depth (Chapter Four). By aligning feature engineering with inferential structure, the thesis shows how PROs, EHR, and high-frequency signals can be synthesized into a unified evidence framework that preserves clinical meaning while scaling to modern data volumes. In this sense, the dissertation operationalizes precision monitoring: classical models ensure clarity and reproducibility; hierarchical Bayes and time-series methods extend this clarity to sparse strata and dynamic settings; together, they deliver decision-relevant, uncertainty-aware inference from integrated, high-dimensional health data [5, 7, 105].

Chapter 6

Conclusion

This thesis advanced a principled, interpretable, and computationally tractable toolkit for learning from complex health data streams across three complementary settings: population surveillance integrating electronic health records (EHR) and patient-reported outcomes (PROs), high-dimensional intracranial electrophysiology, and perioperative pediatric monitoring using processed EEG. Across these settings, the central thread was methodological rigor coupled with interpretability: careful feature construction, likelihoods aligned with data-generating mechanisms, and estimators whose parameters map to clinically meaningful contrasts. Throughout, we combined Frequentist and Bayesian ideas in a pragmatic way—leveraging GEEs for population-averaged effects, hierarchical Bayes for multilevel structure and partial pooling, and state-space/segmentation methods to bridge second-by-second signals with cohort inference [105, 116, 152–154].

Chapter Two analyzes rely on routinely collected PRO sources that are vulnerable to classical observational biases: outcome and exposure misclassification, and residual confounding by unmeasured care-seeking or access variables. Multiple imputation and sensitivity analyzes reduce, but cannot eliminate bias when missingness is not missing at random [69, 149]. In Chapters Three and Four, the cohorts are necessarily small and based on convenience (intracranial recordings from 16 implanted subjects; 20 pediatric surgical cases), which constrains precision for between-subject effects and limits external validity. Hierarchical modeling partially mitigates low subject counts via shrinkage; however, target populations should be interpreted cautiously.

In Chapter Three, the spectral stability analysis used two five-minute windows per day (late morning and late evening) to anchor a diurnal contrast. This design provides clean comparability but omits other circadian states and may miss transient dynamics between anchor times. The FOOOF parameterization depends on algorithmic settings and the 1/f fit domain; while selected to match best practices, different parameterizations could shift the absolute values of aperiodic components [7]. The robust bivariate Student–t likelihood captured heavy tails; however, we modeled only the first-order moments of features; higher-order or cross-frequency structures were not explored.

Chapter Two's macro time-series choices (e.g., ARIMA vs. BSTS) trade interpretability for flexibility [151, 152]. Chapter Three's multilevel models emphasized random intercepts for channels within patients; random slopes (e.g., channel-specific diurnal effects) or spatial priors over electrode geometry could further improve biological fidelity. In Chapter Four, phase segmentation is clinically intuitive but coarse; change point algo-

rithms can be sensitive to penalty choice and minimum segment lengths [154]. Our GEE is correctly interpreted at the population level, but it deliberately avoids subject-specific inferences and assumes correct specification of the mean model; robust (sandwich) SEs guard against misspecified working correlation, not misspecified means [116, 153].

Future work should embed target trial emulation for specific public-health questions (e.g., intervention timing or messaging strategies) atop the same PRO substrate, with negative and positive controls to probe for residual confounding. Dynamic Bayesian models with explicit mechanistic components (e.g., seasonality with change points) can reconcile macro trends with micro-level behavioral shifts, while principled missing-data models allowing MNAR mechanisms (selection or pattern mixture) can bound bias [69, 149, 152]. Multi-site harmonization using common data models and cross-site prior exchange would enable transportable predictions.

Building on the robust, interpretable features, the next steps include (i) extending beyond aperiodic parameters to joint periodic–aperiodic modeling with identifiable priors, (ii) random-slope and spatially structured hierarchical effects that borrow strength across neighboring contacts, and (iii) dynamic state-space models on the feature trajectories to separate tonic from phasic components. Posterior predictive calibration and leave-one-channel/time cross-validation should be reported to adjudicate model choice and quantify prospective stability [105, 106]. Finally, replicating acquisitions, including sleep and task variation, would map stability strata across behavioral states, improving biological generality [7].

Methodologically, Bayesian online change point filters can complement offline PELT by producing real-time probabilities of instability; coupling these with marginal (GEE) calibration curves would retain population interpretability while enabling bedside use. Clinically, multicenter cohorts with standardized PSI acquisition and harmonized artifact control are needed to validate VARI and the effects of the covariates seen here. Richer covariates (drug dosing trajectories, temperature curves, and perfusion parameters) should be integrated through joint models or augmented state-space systems. Subgroup fairness—age, weight, and pathology strata—should be explicitly audited. Prospective studies should test alarm logic that combines macro phase context with micro instability probability and report the impact on clinician behavior and outcomes.

The contributions of this thesis are twofold. Substantively, they provide calibrated summaries of health states across disparate data types: risk contrasts in population surveillance, stability strata in intracranial spectra, and the probability of instability in pediatric anesthesia. Methodologically, they demonstrate how transparent feature engineering, hierarchical modeling, and segmentation/state-space ideas can be assembled into workflows that are both statistically principled and clinically legible [7, 105, 106, 116, 152, 154]. Limitations notwithstanding, the results set the stage for prospective, multicenter evaluations in which interpretable probabilistic outputs inform real decisions—from public messaging to operating-room titration. In doing so, the work argues for a pragmatic synthesis: let modeling choices be driven by the scale at which clinicians act, but disciplined by likelihoods, priors, and diagnostics that ensure validity and reproducibility.

Bibliography

[1] OpenStax. Introduction to behavioral neuroscience. https://openstax.org/books/introduction-behavioral-neuroscience, 2024. Figure 1.31, licensed CC BY 4.0.

[2] M. E. Horn, E. K. Reinke, R. C. Mather, J. D. O'Donnell, and S. Z. George. Electronic health record–integrated approach for collection of patient-reported outcome measures: a retrospective evaluation. *BMC Health Services Research*, 21:626, 2021.

[3] M. J. Bayarri and J. O. Berger. The interplay of bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80, 2004.

[4] Robert Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

[5] All of Us Research Program Investigators. The "all of us" research program. *New England Journal of Medicine*, 381(7):668–676, 2019.

[6] Chisato Imai, Ben Armstrong, Zaid Chalabi, Punam Mangtani, and Masahiro Hashizume. Time series regression model for infectious disease and weather. *Environmental Research*, 142:319–327, 2015.

[7] Thomas Donoghue, Mark Haller, Erik J. Peterson, Priya Varma, Paul Sebastian, Richard Gao, Travis Noto, Robert T. Knight, Anastasia Shestyuk, and Bradley Voytek. Parameterizing neural power spectra into periodic and aperiodic components. *Nature Neuroscience*, 23(12):1655–1665, 2020.

[8] A. H. Hawasli and S. et al. Mofakham. White matter circuits contribute to power of low and high frequency oscillations. *Frontiers in Human Neuroscience*, 9:330, 2016.

[9] G. Schneider, S. Heglmeier, J. Schneider, G. Tempel, and E. F. Kochs. Patient state index (psi) measures depth of sedation in intensive care patients. *Intensive Care Medicine*, 30(2):213–216, 2004.

[10] Michael et al. Soehle. Patient state index (psi) and bis monitoring in pediatric anesthesia. *J. Clin. Monit. Comput.*, 24:123–130, 2010.

[11] Yi Sun, Changwei Wei, Victoria Cui, Meihong Xiu, and Anshi Wu. Electroencephalography: Clinical applications during the perioperative period. *Frontiers in Medicine (Lausanne)*, 7:251, 2020.

[12] Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

[13] Steven L. Scott and Hal R. Varian. Predicting the present with bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, 5(1/2):4–23, 2014.

[14] K.-Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[15] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[16] Rbc Allard. Use of time-series analysis in infectious disease surveillance. *Bulletin of the World Health Organization*, 76(4):327, 1998.

[17] Subhash Kumar Yadav and Yusuf Akhter. Statistical modeling for the prediction of infectious disease dissemination with special reference to covid-19 spread. *Frontiers in public health*, 9:645405, 2021.

[18] Michael R Johnson, Hiten Naik, Wei Siang Chan, Jesse Greiner, Matt Michaleski, Dong Liu, Bruno Silvestre, and Ian P McCarthy. Forecasting ward-level bed requirements to aid pandemic resource planning: Lessons learned and future directions. *Health Care Management Science*, 26(3):477–500, 2023.

[19] Chisato Imai, Ben Armstrong, Zaid Chalabi, Punam Mangtani, and Masahiro Hashizume. Time series regression model for infectious disease and weather. *Environmental research*, 142:319–327, 2015.

[20] Latchezar Tomov, Lyubomir Chervenkov, Dimitrina Georgieva Miteva, Hristiana Batselova, and Tsvetelina Velikova. Applications of time series analysis in epidemiology: Literature review and our experience during covid-19 pandemic. *World Journal of Clinical Cases*, 11(29):6974, 2023.

[21] M Jésus Bayarri and James O Berger. The interplay of bayesian and frequentist analysis. *Statistical Science*, 2004.

[22] Matthias Flor, Michael Weiß, Thomas Selhorst, Christine Müller-Graf, and Matthias Greiner. Comparison of Bayesian and frequentist methods for prevalence estimation under misclassification. *BMC Public Health*, 20(1):1135, December 2020. ISSN 1471-2458. doi: 10.1186/s12889-020-09177-4.

[23] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[24] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[25] Donald B Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, 1987.

[26] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Hoboken, NJ, 2004.

[27] Ian R White, Patrick Royston, and Angela M Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4): 377–399, 2011.

[28] Hongmei Zhu and Joseph G Ibrahim. Bayesian joint modeling of longitudinal and time-to-event data in the presence of missingness: a review. *Statistical Methods in Medical Research*, 30(6):1459–1478, 2021.

[29] Robert M Cronin, Rebecca N Jerome, Brandy Mapes, Regina Andrade, Rebecca Johnston, Jennifer Ayala, David Schlundt, Kemberlee Bonnet, Sunil Kripalani, Kathryn Goggins, et al. Development of the initial surveys for the all of us research program. *Epidemiology*, 30(4):597–608, 2019.

[30] All of Us Research Program Investigators. The "all of us" research program. *New England Journal of Medicine*, 381(7):668–676, 2019.

[31] Tribal Collaboration Working Group et al. All of us research program advisory panel. *Considerations for Meaningful Collaboration With Tribal Populations*, 2018.

[32] R R Core Team et al. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, 2013.

[33] Martyn Plummer et al. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–10. Vienna, Austria, 2003.

[34] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76 (1):1–32, 2017.

[35] OpenBUGS Project. *OpenBUGS: Open Source Implementation of the BUGS Language*, 2023. URL http://www.openbugs.net/.

[36] Paul-Christian Bürkner. brms: An r package for bayesian multilevel models using stan, 2017. URL https://doi.org/10.18637/jss.v080.i01.

[37] Martyn Plummer. *rjags: Bayesian Graphical Models Using MCMC*, 2023. URL https://CRAN.R-project.org/package=rjags.

[38] Stan Development Team. *RStan: the R interface to Stan*, 2023. URL https://mc-stan.org/users/interfaces/rstan.

[39] Dominique Makowski, Mattan S Ben-Shachar, SH Chen, and Daniel Lüdecke. bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, 4(40):1541, 2019.

[40] John K Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press, 2 edition, 2015.

[41] John K Kruschke. Rejecting or accepting parameter values in bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280, 2018.

[42] Eric-Jan Wagenmakers, Tom Lodewyckx, Himanshu Kuriyal, and Raoul Grasman. Bayesian hypothesis testing for psychologists: A tutorial on the savage–dickey method. *Cognitive Psychology*, 60(3):158–189, 2010.

[43] Heinz Schmidli, Sandro Gsteiger, Satrajit Roychoudhury, Anthony O'Hagan, David Spiegelhalter, and Beat Neuenschwander. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032, 2014.

[44] Isabelle Albert, Sophie Donnet, Chantal Guihenneuc-Jouyaux, Samantha Low-Choy, Kerrie Mengersen, and Judith Rousseau. Combining expert opinions in prior elicitation (Pkg: p503-546). *Bayesian Analysis*, 7(3):503–532, 2012. doi: 10.1214/12-BA717.

[45] Marta Soares. Recommendations on the use of structured expert elicitation protocols for healthcare decision making: A good practices report of an ispor task force. *Value in Health*, 2024.

[46] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of mcmc. *Bayesian Analysis*, 16(2):667–718, 2021.

[47] Annette J Dobson and Adrian G Barnett. *An introduction to generalized linear models*. Chapman and Hall/CRC, 2018.

[48] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3):370–384, 1972.

[49] Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

[50] Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1): 1593–1623, 2014.

[51] Xinliang Liu, Yuxi Long, Christine Greenhalgh, Sarah Steeg, Jack Wilkinson, Hao Li, Arpana Verma, and Angela Spencer. A systematic review and meta-analysis of risk factors associated with healthcare-associated infections among hospitalized patients in chinese general hospitals from 2001 to2022. *Journal of Hospital Infection*, 135:37–49, 2023.

[52] Priya Ranganathan, Rakesh Aggarwal, and C S Pramesh. Common pitfalls in statistical analysis: Odds versus risk. *Perspectives in Clinical Research*, 6(4):222–224, 2015. doi: 10.4103/2229-3485.167092. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4640017/.

[53] Huw T O Davies, Iain K Crombie, and Mohammad Tavakoli. When can odds ratios mislead? *BMJ*, 316(7136):989–991, 1998. doi: 10.1136/bmj.316.7136.989. URL https://doi.org/10.1136/bmj.316.7136.989.

[54] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[55] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition, 2009. ISBN 9780387848574.

[56] Ruben Dezeure, Peter Bühlmann, Lukas Meier, and Nicolai Meinshausen. High-dimensional inference: confidence intervals, p-values and r-software hdi. *Statistical Science*, 30(4):533–558, 2015. doi: 10.1214/15-STS527.

[57] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010. doi: 10.1111/j.1467-9868.2010.00740.x.

[58] Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[59] Francisco J Samaniego. *A comparison of the Bayesian and frequentist approaches to estimation*, volume 24. Springer, 2010.

[60] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[61] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S*. Springer, fourth edition, 2002.

[62] Kenneth P. Burnham and David R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, 2002.

[63] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27:1413–1432, 2017.

[64] Jayanta K. Ghosh, Mohan Delampady, and Tapas Samanta. *An Introduction to Bayesian Analysis: Theory and Methods*. Springer, New York, 2006. ISBN 9780387354334.

[65] Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[66] James A Hanley, Abdissa Negassa, Michael D deB Edwardes, and Janet E Forrester. Statistical analysis of correlated data using generalized estimating equations: an orientation. *American journal of epidemiology*, 157(4):364–375, 2003.

[67] James W. Hardin and Joseph M. Hilbe. *Generalized Estimating Equations*. Chapman & Hall/CRC, 2002.

[68] X. Wang and J. Smith. Covariance estimation for clustered data. *Journal of Statistical Methods*, 45:123–135, 2019.

[69] Stef van Buuren. *Flexible Imputation of Missing Data*. CRC Press, 2018.

[70] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011.

[71] Paul L. Nunez and Ramesh Srinivasan. *Electric Fields of the Brain: The Neurophysics of EEG*. Oxford University Press, 2nd edition, 2006.

[72] György Buzsáki, Costas A. Anastassiou, and Christof Koch. The origin of extracellular fields and currents – eeg, ecog, lfp and spikes. *Nature Reviews Neuroscience*, 13:407–420, 2012. doi: 10.1038/nrn3241.

[73] Josef Parvizi and Sabine Kastner. Promises and limitations of human intracranial electroencephalography. *Nature Neuroscience*, 21:474–483, 2018. doi: 10.1038/s41593-018-0108-2.

[74] Jean-Philippe Lachaux, Nikolai Axmacher, Florian Mormann, Eric Halgren, and Nathan E. Crone. High-frequency neural activity and human cognition: Past, present and possible future of intracranial eeg research. *Progress in Neurobiology*, 98(3):279–301, 2012. doi: 10.1016/j.pneurobio.2012.06.008.

[75] Karim Jerbi, Tomas Ossandon, Carlos M. Hamamé, and et al. Task-related gamma-band dynamics from an intracerebral perspective: Review and implications for surface eeg and meg. *Human Brain Mapping*, 30(6):1758–1771, 2009. doi: 10.1002/hbm.20750.

[76] Kai J. Miller, Stavros P. Zanos, Eberhard E. Fetz, Marcel den Nijs, and Jeffrey G. Ojemann. Decoupling the cortical power spectrum reveals real-time representation of individual finger movements in humans. *Journal of Neuroscience*, 29(10):3132–3137, 2009. doi: 10.1523/JNEUROSCI.5506-08.2009.

[77] Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar. Moving to a world beyond "p < 0.05". *The American Statistician*, 73(sup1):1–19, 2019. doi: 10.1080/00031305.2019.1583913.

[78] B. L. Welch. The generalization of "student's" problem when several different population variances are involved. *Biometrika*, 34(1–2):28–35, 1947.

[79] Peter D. Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, 1967. doi: 10.1109/TAU.1967.1161901.

[80] Donald B. Percival and Andrew T. Walden. *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques*. Cambridge University Press, 1993.

[81] Fredric J. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978. doi: 10.1109/PROC.1978.10837.

[82] Petre Stoica and Randolph L. Moses. *Spectral Analysis of Signals*. Prentice Hall, 2005.

[83] José C. Pinheiro and Douglas M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, 2000.

[84] Ian T. Jolliffe. *Principal Component Analysis*. Springer, 2nd edition, 2002.

[85] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000. doi: 10.1016/S0893-6080(00)00026-5.

[86] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999. doi: 10.1038/44565.

[87] Scott R. Cole and Bradley Voytek. Brain oscillations and the importance of waveform shape. *Trends in Cognitive Sciences*, 21(2):137–149, 2017. doi: 10.1016/j.tics.2016.12.008.

[88] Matar Haller, Thomas Donoghue, Evan J. Peterson, and et al. Parameterizing neural power spectra. *bioRxiv*, page 299859, 2018. doi: 10.1101/299859.

[89] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200): 675–701, 1937.

[90] M. G. Kendall and B. Babington Smith. The problem of *m* rankings. *Annals of Mathematical Statistics*, 10(3):275–287, 1939.

[91] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

[92] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. doi: 10.2307/2529310.

[93] Alboukadel Kassambara. Friedman test effect size (kendall's w) — `friedman_effsize`. https://rpkgs.datanovia.com/rstatix/reference/friedman_effsize.html, 2024. Accessed 2025-08-27.

[94] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[95] Paul-Christian Bürkner. brms: An R package for bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28, 2017.

[96] D. Kumral et al. Relationship between regional white matter hyperintensities and changes in alpha oscillations in older adults. *Neurobiology of Aging*, 110:123–134, 2022.

[97] S. Makeig and M. Inlow. Lapse in alertness: coherence of fluctuations in performance and eeg spectrum. *Electroencephalography and Clinical Neurophysiology*, 86:23–35, 1993.

[98] G. Buzsáki, C. A. Anastassiou, and C. Koch. The origin of extracellular fields and currents—eeg, ecog, lfp and spikes. *Nature Reviews Neuroscience*, 13:407–420, 2012.

[99] H. G. Kwon, W. M. Byun, S. H. Ahn, S. M. Son, and S. H. Jang. The anatomical characteristics of the stria terminalis in the human brain: a diffusion tensor tractography study. *Neuroscience Letters*, 500:99–102, 2011. doi: 10.1016/j.neulet.2011.06.026.

[100] M. Catani and M. Thiebaut de Schotten. A diffusion tensor imaging tractography atlas for virtual in vivo dissections. *Cortex*, 44:1105–1132, 2006.

[101] John K. Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press, 2nd edition, 2014.

[102] Paul-Christian Bürkner. brms: An R package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1):1–28, 2017.

[103] Andrew Gelman, Aleks Jakulin, Maria Giulia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008. doi: 10.1214/08-AOAS191.

[104] Daniel Lewandowski, Dorota Kurowicka, and Harry Joe. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001, 2009. doi: 10.1016/j.jmva.2009.04.008.

[105] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, 3rd edition, 2013.

[106] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5): 1413–1432, 2017.

[107] Peter McCullagh and John A. Nelder. *Generalized Linear Models*. Chapman and Hall/CRC, 2nd edition, 1989.

[108] Jose M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley, 1994.

[109] Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. In *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC, 2017. arXiv:1701.02434.

[110] Theodore W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 3rd edition, 2003.

[111] Alvin C. Rencher and William F. Christensen. *Methods of Multivariate Analysis*. Wiley, 2nd edition, 2002.

[112] José C. Pinheiro and Douglas M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, 2000.

[113] Robert Oostenveld, Pascal Fries, Eric Maris, and Jan-Mathijs Schoffelen. Fieldtrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, page 156869, 2011. doi: 10.1155/2011/156869.

[114] *R: A Language and Environment for Statistical Computing*. R Core Team, Vienna, Austria, 2024. URL https://www.R-project.org/.

[115] Mitzi Morris, Katherine Wheeler-Martin, Dan Simpson, Stephen J. Mooney, Andrew Gelman, and Charles DiMaggio. Bayesian hierarchical spatial models: Implementing the besag–york–mollié (bym) model in stan. Case study, Stan Development Team, 2019. URL https://mc-stan.org/users/documentation/case-studies/icar_stan.html.

[116] Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986. doi: 10.1093/biomet/73.1.13.

[117] Rebecca Killick and Idris A. Eckley. changepoint: An r package for changepoint analysis. *Journal of Statistical Software*, 58(3), 2014. doi: 10.18637/jss.v058.i03.

[118] Steven L. Scott and Hal R. Varian. Predicting the present with Bayesian structural time series. *IJMMNO*, 5(1/2):4, 2014. ISSN 2040-3607, 2040-3615. doi: 10.1504/IJMMNO.2014.059942. URL http://www.inderscience.com/link.php?id=59942.

[119] Gerhard Schneider, Susanne Heglmeier, Jürgen Schneider, Gunter Tempel, and Eberhard F. Kochs. Patient State Index (PSI) measures depth of sedation in intensive care patients. *Intensive Care Med*, 30(2):213–216, February 2004. ISSN 0342-4642, 1432-1238. doi: 10.1007/s00134-003-2092-5. URL http://link.springer.com/10.1007/s00134-003-2092-5.

[120] David Drover and H.R. (Rick) Ortega. Patient state index. *Best Practice & Research Clinical Anaesthesiology*, 20(1):121–128, March 2006. ISSN 15216896. doi: 10.1016/j.bpa.2005.07.008. URL https://linkinghub.elsevier.com/retrieve/pii/S152168960500056X.

[121] Michael Soehle and et al. Patient state index (psi) and bis monitoring in pediatric anesthesia. *Journal of Clinical Monitoring and Computing*, 24:123–130, 2010.

[122] W.G. Muhlhofer, R. Zak, T. Kamal, B. Rizvi, L.P. Sands, M. Yuan, X. Zhang, and J.M. Leung. Burst-suppression ratio underestimates absolute duration of electroencephalogram suppression compared with visual analysis of intraoperative electroencephalogram. *British Journal of Anaesthesia*, 118(5):755–761, May 2017. ISSN 00070912. doi: 10.1093/bja/aex054. URL https://linkinghub.elsevier.com/retrieve/pii/S0007091217313363.

[123] Yi Sun, Changwei Wei, Victoria Cui, Meihong Xiu, and Anshi Wu. Electroencephalography: Clinical Applications During the Perioperative Period. *Front Med (Lausanne)*, 7:251, 2020. ISSN 2296-858X. doi: 10.3389/fmed.2020.00251.

[124] Z. Gao, J. Zhang, X. Wang, et al. A retrospective study of electroencephalography burst suppression in children undergoing general anesthesia. *Pediatric Investigation*, 5(4):271–276, 2021. doi: 10.1002/ped4.12287.

[125] Masafumi Idei, Yusuke Seino, Nobuo Sato, Takuya Yoshida, Yumi Saishu, Kimiya Fukui, Masahiro Iwabuchi, Junya Ishikawa, Kei Ota, Daigo Kamei, Masashi Nakagawa, and Takeshi Nomura. Validation of the patient State Index for monitoring sedation state in critically ill patients: a prospective observational study. *J Clin Monit Comput*, 37(1):147–154, February 2023. ISSN 1387-1307, 1573-2614. doi: 10.1007/s10877-022-00871-9. URL https://link.springer.com/10.1007/s10877-022-00871-9.

[126] Zaccaria Ricci, Chiara Robino, Paolo Rufini, Silvia Cumbo, Sara Cavallini, Lorenzo Gobbi, Agata Brocchi, Paola Serio, and Stefano Romagnoli. Monitoring anesthesia depth with patient state index during pediatric surgery. *Pediatric Anesthesia*, 33(10):855–861, October 2023. ISSN 1155-5645, 1460-9592. doi: 10.1111/pan.14711. URL https://onlinelibrary.wiley.com/doi/10.1111/pan.14711.

[127] R. Killick, P. Fearnhead, and I. A. Eckley. Optimal Detection of Changepoints With a Linear Computational Cost. *Journal of the American Statistical Association*, 107(500):1590–1598, December 2012. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2012.737745. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2012.737745.

[128] Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, February 2020. ISSN 01651684. doi: 10.1016/j.sigpro.2019.107299. URL https://linkinghub.elsevier.com/retrieve/pii/S0165168419303494.

[129] Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. Inferring causal impact using bayesian structural time–series models. *Annals of Applied Statistics*, 9(1):247–274, 2015. doi: 10.1214/14-AOAS788.

[130] Norman L. Johnson, Adrienne W. Kemp, and Samuel Kotz. *Univariate Discrete Distributions*. Wiley, Hoboken, NJ, 3 edition, 2005.

[131] Wei Pan. Akaike's information criterion in generalized estimating equations. *Biometrics*, 57(1):120–125, 2001.

[132] J Diggle Peter, Heagerty Patrick, Liang Kung-Yee, and L Zeger Scott. Analysis of longitudinal data. *Oxford Statistical Science Series. OUP Oxford*, 2002.

[133] L. Yu. Vostrikova. Detection of the disorder in stochastic processes. *Theory of Probability and Its Applications*, 26(2):379–388, 1981.

[134] Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. *Annals of Statistics*, 42(6):2243–2281, 2014. doi: 10.1214/14-AOS1245.

[135] Jesse M. Ehrenfeld. Anesthetic Techniques: General, Sedation, MAC. In Jesse M. Ehrenfeld, Richard D. Urman, and B. Scott Segal, editors, *Anesthesia Student Survival Guide*, pages 217–233. Springer International Publishing, Cham, 2022. ISBN 978-3-030-98674-2 978-3-030-98675-9. doi: 10.1007/978-3-030-98675-9_12. URL https://link.springer.com/10.1007/978-3-030-98675-9_12.

[136] Ricardo A. Maronna and R. Douglas Martin and Victor J. Yohai and Matías Salibián-Barrera. *Robust Statistics - Theory and Methods (with R) Second Edition*. Wiley, 2019. ISBN 978-1-119-21465-6.

[137] Alessio Farcomeni and Laura Ventura. An overview of robust methods in medical research. *Stat Methods Med Res*, 21(2):111–133, April 2012. ISSN 0962-2802, 1477-0334. doi: 10.1177/0962280210385865. URL http://journals.sagepub.com/doi/10.1177/0962280210385865.

[138] Ricardo A. Maronna, R. Douglas Martin, and Víctor J. Yohai. *Robust statistics: theory and methods*. Wiley series in probability and statistics. J. Wiley, Chichester (GB), 2006. ISBN 978-0-470-01092-1.

[139] Karl Pearson. X. Contributions to the mathematical theory of evolution.—II. Skew variation in homogeneous material. *Phil. Trans. R. Soc. Lond. A*, 186:343–414, December 1895. ISSN 0264-3820, 2053-9231. doi: 10.1098/rsta.1895.0010. URL https://royalsocietypublishing.org/doi/10.1098/rsta.1895.0010.

[140] Mohamed Elfil and Ahmed Negida. Sampling methods in Clinical Research; an Educational Review. *Emerg (Tehran)*, 5(1):e52, 2017. ISSN 2345-4563.

[141] B. Efron. Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist.*, 7(1), January 1979. ISSN 0090-5364. doi: 10.1214/aos/1176344552. URL https://projecteuclid.org/journals/annals-of-statistics/volume-7/issue-1/Bootstrap-Methods-Another-Look-at-the-Jackknife/10.1214/aos/1176344552.full.

[142] Bradley Efron. Bayesian inference and the parametric bootstrap. *Ann Appl Stat*, 6 (4):1971–1997, October 2012. ISSN 1932-6157. doi: 10.1214/12-AOAS571.

[143] Julien Bert and David Sarrut. Monte Carlo simulations for medical and biomedical applications. In *Biomedical Image Synthesis and Simulation*, pages 23–53. Elsevier, 2022. ISBN 978-0-12-824349-7. doi: 10.1016/B978-0-12-824349-7.00010-4. URL https://linkinghub.elsevier.com/retrieve/pii/B9780128243497000104.

[144] Norman Lloyd Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous univariate distributions*. Wiley series in probability and mathematical statistics. Wiley, New York, 2nd ed edition, 1994. ISBN 978-0-471-58495-7 978-0-471-58494-0.

[145] Nathan P. Lemoine. Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. *Oikos*, 128(7):912–928, July 2019. ISSN 0030-1299, 1600-0706. doi: 10.1111/oik.05985. URL https://onlinelibrary.wiley.com/doi/10.1111/oik.05985.

[146] Raydonal Ospina and Silvia L.P. Ferrari. A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, 56(6):1609–1623, June 2012. ISSN 01679473. doi: 10.1016/j.csda.2011.10.005. URL https://linkinghub.elsevier.com/retrieve/pii/S0167947311003628.

[147] M. J. Bayarri and J. O. Berger. The interplay of bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80, 2004. doi: 10.1214/088342304000000116.

[148] Francisco J. Samaniego. *A Comparison of the Bayesian and Frequentist Approaches to Estimation*. Springer Series in Statistics. Springer, 2010. ISBN 978-1441968695. doi: 10.1007/978-1-4419-6869-5.

[149] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, 1987. ISBN 978-0471655749. doi: 10.1002/9780470316696.

[150] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. doi: 10.18637/jss.v045.i03.

[151] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 5 edition, 2015. ISBN 978-1118675021.

[152] Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics*, 9(1):247–274, 2015. doi: 10.1214/14-AOAS788.

[153] James A. Hanley, Abdissa Negassa, Michael D. Edwardes, and Janet E. Forrester. Statistical analysis of correlated data using generalized estimating equations: an orientation. *American Journal of Epidemiology*, 157(4):364–375, 2003. doi: 10.1093/aje/kwf215.

[154] Rebecca Killick, Paul Fearnhead, and Idris A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012. doi: 10.1080/01621459.2012.737745.