The New England Journal of Statistics in Data Science Volume 0, 1–23 (2026)
DOI: https://doi.org/10.51387/26-NEJSDS104

Efficacy Analysis in Clinical Trials: A Comprehensive Review of Statistical and Machine Learning Approaches

Dhrubajyoti GHOSH* and Samhita PAL[1]

Abstract

Efficacy testing is a cornerstone of clinical trials, ensuring that medical interventions achieve their intended therapeutic effects. Over the decades, a wide range of statistical methodologies have been developed to address the complexities of clinical trial data, including parametric, nonparametric, Bayesian, and machine learning approaches. Parametric methods, such as t-tests, ANOVA, and LMMs, have traditionally been the foundation of efficacy testing due to their efficiency under well-defined assumptions. Nonparametric techniques, including the Friedman test, Brunner-Munzel test, and modern extensions like nparLD, have emerged as robust alternatives, particularly for skewed, ordinal, or non-normal data. Bayesian methodologies have enabled the incorporation of prior information and uncertainty quantification, while machine learning techniques, such as deep learning and reinforcement learning, are revolutionizing trial designs and outcome predictions. Despite these advancements, significant gaps remain, including challenges in handling high-dimensional data, missingness, and ensuring equitable efficacy testing across diverse populations. This review provides a comprehensive overview of these statistical methods, highlighting their applications, strengths, limitations, and future directions. By bridging traditional statistical frameworks with modern computational techniques, the field can continue to advance toward more reliable and personalized clinical trial methodologies.

AMS 2000 subject classifications: Primary 62P10, 62F03; secondary 62G10.
keywords and phrases: Efficacy testing, Longitudinal, Cross-sectional, Clinical trials, Parametric methods, Nonparametric methods, Bayesian methods, Machine learning, Deep learning.

CONTENTS

1. INTRODUCTION

Efficacy testing is a fundamental aspect of clinical trials, assessing whether a medical intervention produces the

arXiv: 2511.04903
*Corresponding author.
[1]These authors contributed equally as first authors.

intended therapeutic effect under ideal or controlled conditions. It serves as the cornerstone for determining the effectiveness of drugs, treatments, and devices, guiding regulatory approvals and informing clinical practice. Statistical methods are pivotal in this process, ensuring that efficacy conclusions are robust, reproducible, and scientifically valid. Over the years, advancements in statistical methodologies have enabled researchers to tackle the complexities inherent in efficacy testing, particularly as datasets have become more intricate and multidimensional.

Parametric methods have traditionally dominated efficacy testing in clinical trials due to their simplicity and statistical efficiency. Classical approaches, such as t-tests and analysis of variance (ANOVA), remain staples for comparing treatment effects across groups [132, 42]. For longitudinal and repeated measures data, linear mixed-effects models (LMMs) have become the method of choice, enabling the modeling of individual patient trajectories while accounting for variability across and within subjects [81, 145]. However, parametric methods rely heavily on assumptions of normality and homogeneity of variance, which are often violated in real-world clinical trials, particularly with heterogeneous or skewed data.

To address these challenges, nonparametric methods have emerged as robust alternatives for efficacy testing. These methods, including the Friedman test [45], Brunner-Munzel test [15], and modern tools like nparLD [98], provide greater flexibility by relaxing distributional assumptions. Nonparametric approaches are particularly suited for ordinal outcomes, non-normal data, and small sample sizes, making them invaluable in certain therapeutic areas and rare disease studies. Additionally, recent advancements, such as the Longitudinal Rank Sum Test (LRST) [158, 49], have further extended the utility of nonparametric methods to longitudinal settings, enabling the analysis of time-dependent efficacy outcomes.

Bayesian methods have also gained prominence in efficacy testing by allowing researchers to incorporate prior knowledge and account for uncertainty in treatment effects. Hierarchical Bayesian models enable the pooling of information across subgroups, while Bayesian nonparametric approaches, such as Gaussian processes, provide flexible tools for modeling complex, nonlinear relationships in longitudinal data [48, 111]. These methods are particularly advantageous in adaptive trial designs, where interim results inform subsequent decisions, and in personalized medicine, where individual-level predictions are critical.

In recent years, machine learning (ML) techniques have revolutionized efficacy testing by addressing challenges associated with high-dimensional, unstructured, and multimodal datasets. Methods such as deep learning, reinforcement learning, and federated learning have been applied to optimize trial designs [54], predict treatment outcomes [164], and improve dosing regimens [96]. Despite their promise, these approaches face challenges related to interpretability, generalizability, and integration with traditional statistical frameworks.

The choice of analytical framework in efficacy testing depends on the underlying research objective and the structure of the clinical data. Broadly, these objectives can be grouped into four categories: estimating and comparing treatment effects under defined assumptions, assessing relative efficacy through distributional comparisons with minimal assumptions, quantifying uncertainty while incorporating prior knowledge, and predicting individual-level outcomes or response patterns in complex data. Parametric methods are most appropriate when the goal is to estimate average treatment differences or model temporal trajectories under well-specified distributional assumptions. Nonparametric approaches are valuable when the data are skewed, ordinal, or heterogeneous, allowing robust comparisons that remain valid under weaker conditions. Bayesian models provide a flexible probabilistic framework that integrates prior evidence with observed data, supporting hierarchical and adaptive trial designs. Finally, machine learning techniques, ranging from deep neural architectures to reinforcement learning, are designed for predictive and exploratory analyses, capable of capturing complex, nonlinear dependencies in high-dimensional or multimodal datasets. By explicitly linking these analytical families to their corresponding research objectives, this review aims to guide researchers in selecting suitable methodologies for efficacy analysis across diverse clinical settings.

Missing data are almost unavoidable in clinical and biomedical research, and the assumptions made about how data become missing have direct consequences for statistical validity. The standard taxonomy distinguishes among three mechanisms [86, 94]. Data are missing completely at random (MCAR) when the probability of missingness is unrelated to both observed and unobserved outcomes; complete-case analysis remains unbiased but inefficient in this setting. Data are missing at random (MAR) when missingness depends only on observed variables; likelihood-based estimation and multiple imputation can then provide valid inference if the model for the observed data is correctly specified. Data are missing not at random (MNAR) when missingness depends on unobserved outcomes, leading to potential bias even under correct modeling of the observed data. In such cases, specialized models such as selection models, pattern-mixture models, or shared-parameter frameworks must explicitly characterize the missingness mechanism [26]. Because these assumptions underpin all major statistical paradigms, the subsequent sections highlight how each methodological family handles missingness under these three mechanisms.

This review provides a comprehensive examination of statistical methods for efficacy testing in clinical trials, categorizing them into parametric, nonparametric, Bayesian, and machine learning approaches. By highlighting the strengths, limitations, and applications of these methodologies, the paper aims to guide researchers in selecting appropriate tools

for efficacy analysis in diverse clinical settings. Furthermore, the review explores recent methodological advancements, emphasizing their potential to address emerging challenges in clinical trial research. We have explored parametric tests in Section 2, nonparametric tests in Section 3, Bayesian methods in Section 4 and some popular Deep Learning based methodologies in Section 5.

2. PARAMETRIC TESTS

Parametric tests are among the most widely used statistical methods in clinical trials, owing to their simplicity, efficiency, and strong inferential properties under well-defined assumptions. Parametric approaches are most suitable when the objective is to estimate mean treatment differences or model trajectories under defined distributional assumptions. These methods rely on specific distributional assumptions, such as normality of the data and homogeneity of variances, making them particularly suited for continuous and normally distributed outcomes. Commonly applied parametric tests include the t-test, analysis of variance (ANOVA), and analysis of covariance (ANCOVA), each tailored for comparing means across groups or conditions under varying experimental designs. In many clinical trials, efficacy is evaluated based on the change from baseline to the final endpoint, where cross-sectional parametric tests like the t-test or ANCOVA can be used to compare treatment groups [146, 106]. For longitudinal or repeated measures data, LMMs extend the scope of parametric analysis by incorporating random effects to account for within-subject correlations and handling missing data under the missing-at-random (MAR) framework [81, 145]. While parametric methods are robust under their assumptions, their utility can be limited when these assumptions are violated, such as in the presence of skewed distributions or heterogeneous variances, highlighting the need for alternative approaches in certain clinical trial settings.

2.1 Cross-Sectional Parametric Tests

Efficacy testing in clinical trials often begins with the comparison of treatment groups at a single time point, where parametric methods provide powerful and interpretable statistical tools. These methods, including t-tests, analysis of variance (ANOVA), and analysis of covariance (ANCOVA), leverage assumptions about data distribution to efficiently estimate treatment effects. The following section discusses these cross-sectional parametric tests, their applications in clinical research, and their advantages and limitations in different trial settings.

2.1.1 t-Tests

The t-test, introduced by William Sealy Gosset in 1908 under the pseudonym "Student," is a parametric statistical test designed to compare the means of two groups under specific assumptions [132]. The standard t-test, also called the independent t-test, evaluates the difference between two independent group means. The test statistic is calculated as: $t = \left(\bar{X}_1 - \bar{X}_2\right) / \sqrt{s_p^2(n_1^{-1} + n_2^{-1})}$, where $\bar{X}_1$ and $\bar{X}_2$ are the sample means, n_1 and n_2 are the sample sizes, and $s_p^2 = \left((n_1-1)s_1^2 + (n_2-1)s_2^2\right) / (n_1+n_2-2)$ is the pooled variance. Here, s_1^2 and s_2^2 represent the variances of the two groups. The paired t-test, a variation of the standard t-test, is used for dependent or matched samples, analyzing the mean differences between paired observations using the formula: $t = \bar{d} / (s_d/\sqrt{n})$, where $\bar{d}$ is the mean of the differences, s_d is the standard deviation of the differences, and n is the number of paired observations. Another important extension, Welch's t-test, modifies the standard t-test to account for unequal variances and sample sizes between groups, with its test statistic computed as: $t = \left(\bar{X}_1 - \bar{X}_2\right) / \sqrt{(s_1^2/n_1) + (s_2^2/n_2)}$. These methods, which rely on the assumption of normally distributed data, have become foundational tools for statistical inference in clinical trials and experimental research [150].

T-tests are widely recognized for their simplicity, computational efficiency, and effectiveness in comparing means, particularly in small to moderate-sized datasets (Cf. Table 1). The paired t-test enhances statistical power by focusing on within-subject differences, while Welch's t-test provides a robust alternative for scenarios involving unequal variances or sample sizes. However, these methods rely on assumptions of normality and are sensitive to outliers, which can compromise their validity in non-normal or heavily skewed data. Additionally, t-tests are limited to mean-based comparisons, rendering them unsuitable for ordinal data or high-dimensional datasets that require more sophisticated statistical approaches, such as mixed-effects models or machine learning techniques.

2.1.2 Variance Analysis Methods

Analysis of Variance (ANOVA), introduced by Fisher in 1918, is a parametric method for comparing the means of multiple groups. It partitions total variation into between-group variation (SS_{between}) and within-group variation (SS_{within}), with the F-statistic calculated as $F = SS_{\text{between}}(N-k)/SS_{\text{within}}(k-1)$ where k is the number of groups and N is the total number of observations. Under the null hypothesis (H_0), the F-statistic follows an F-distribution with $k-1$ and $N-k$ degrees of freedom. ANOVA assumes normality within groups, homoscedasticity, and independence of observations. Extensions like repeated measures and two-way ANOVA have been developed for longitudinal and factorial designs. While ANOVA controls Type I error better than multiple pairwise t-tests, violations of its assumptions can lead to misleading results. Nonparametric alternatives like the Kruskal-Wallis test address these limitations. ANOVA is also less powerful with unequal group sizes.

Analysis of Covariance (ANCOVA), introduced by Fisher in 1925, combines ANOVA and linear regression to compare group means while adjusting for one or more covariates. ANCOVA is particularly useful in clinical trials to

Table 1. Applications of the t-Tests in Clinical Trials.

Study	Application	Therapeutic Area
[11]	Evaluation of pre- and post-treatment differences in blood pressure levels	Cardiovascular Research
[109]	Assessment of the impact of antihypertensive therapies on systolic and diastolic blood pressure in hypertensive patients	Hypertension
[47]	Analysis of Metformin's impact on glycemic control using independent t-tests to compare pre- and post-intervention HbA1c levels	Diabetes
[44]	Analysis of the efficacy of dietary and behavioral therapies in weight-loss interventions	Weight-Loss Interventions
[165]	Evaluation of psychometric properties in mental health research	Mental Health
[147]	Evaluation of cognitive behavioral therapy's impact on depression scores in mental health interventions	Mental Health
[40]	Examination of the effects of angiotensin-converting enzyme inhibitors on cardiac function	Cardiology
[117]	Comparison of drug efficacy for treatments with differing variance profiles	General Healthcare
[131]	Systematic reviews to synthesize data on pre- and post-treatment changes in diagnostic accuracy studies	Meta-analytical Frameworks
[38]	Application of Welch's t-test in contingency table analyses for robustness in healthcare data	Healthcare
[110]	Exploration of Welch's t-test in pharmacokinetics to address unequal variances due to population heterogeneity	Pharmacokinetics
[31]	Highlighted effectiveness of Welch's t-test for small-sample comparisons in clinical studies	Clinical Research
[103]	Assessment of the efficacy of Liraglutide in reducing body weight through paired analysis	Weight Management
[9]	Examination of Budesonide-Formoterol therapy in improving lung function parameters through paired t-tests	Pulmonology
[130]	Optimization of photon-counting CT for lung density quantifications using paired analysis	Radiology

control for baseline imbalances (Cf. Table 2). The model is $Y_{ij} = \mu + \alpha_i + \beta X_{ij} + \epsilon_{ij}$ where Y_{ij} is the outcome, X_{ij} is the covariate, and β is the regression coefficient for the covariate. ANCOVA adjusts the outcome by removing the effect of the covariate, enabling a comparison of group means for a common covariate value. It assumes linearity between the covariate and the outcome, homogeneity of regression slopes, and normality and homoscedasticity of residuals. ANCOVA has been extended to handle multiple covariates and more complex designs. However, violations of assumptions or failure to account for covariate-treatment interactions can lead to biased results.

2.2 Longitudinal Parametric Tests

While cross-sectional parametric tests such as t-tests and ANOVA are widely used in clinical trials, they are limited to analyzing treatment effects at a single time point. However, many clinical studies involve repeated measurements of patient outcomes over time, requiring statistical methods that account for within-subject correlations and time-dependent variations. Longitudinal parametric models, such as Linear Mixed-Effects Models (LMMs) and Generalized Estimating Equations (GEEs), extend traditional parametric frameworks to capture temporal dynamics and patient-specific variability. The following section explores these methods, highlighting their applications in efficacy testing and their ability to handle missing data and complex correlation structures.

2.2.1 Linear Mixed Effects Model

LMMs, introduced by [61], extend linear regression by incorporating both fixed and random effects, making them particularly suitable for analyzing hierarchical or longitudinal data, such as repeated measures in clinical trials. Fixed effects represent population-level relationships, while random effects capture subject-specific variability, allowing LMMs to effectively model within-subject or within-cluster correlations. The general model is expressed as: $Y = X\beta + Zb + \epsilon$, where $X\beta$ denotes the fixed effects, Zb represents the random effects ($b \sim N(0, G)$), and $\epsilon \sim N(0, R)$ accounts for residual errors. One of the significant strengths of LMMs is their ability to handle missing data under the Missing at Random (MAR) assumption, where the probability of missingness depends only on observed data. Unlike traditional methods that exclude incomplete cases, LMMs use all available data through maximum likelihood or restricted maximum likelihood estimation, reducing bias and improving efficiency. Additionally, LMMs accommodate complex covariance structures, such as autoregressive or unstructured correlations, providing flexibility in modeling dependencies.

LMMs, while versatile, have limitations. They assume linear relationships, normality of random effects and residuals,

Table 2. Applications of ANOVA and ANCOVA in Clinical Trials.

Study	Method	Application	Therapeutic Area
[99]	ANOVA	Multivariate studies of psychological interventions for chronic pain management	Chronic Pain Management
[73]	ANCOVA	Adjustment for tumor size at baseline in oncology trials to compare treatment efficacy	Oncology
[105]	ANOVA	Comparison of the effects of beta-blockers across different dosages	Cardiovascular Research
[155]	ANCOVA	Longitudinal studies to address baseline imbalances in randomized controlled trials	Randomized Controlled Trials
[122]	ANOVA	Comparison of cognitive behavioral therapies across different durations of intervention	Mental Health
[115]	ANCOVA	A 24-week, double-blind, placebo-controlled trial of donepezil in Alzheimer's disease	Alzheimer's Disease
[104]	ANOVA	Extension to mixed models for longitudinal analyses in oncology trials	Oncology
[146]	ANCOVA	Adjusting for percentage change from baseline in pain management studies, avoiding biased estimates	Pain Management
[10]	ANOVA	Assessment of musculoskeletal burden differences across patient subgroups in observational studies	Musculoskeletal Disorders
[106]	ANCOVA	Cardiovascular trials to adjust for baseline blood pressure and cholesterol levels	Cardiovascular Research
[14]	ANCOVA	Survival analysis to adjust for clinical staging in oncology trials	Oncology
[91]	ANOVA	Use of repeated measures to evaluate temporal effects of pain relief medications in drug trials	Pain Management
[56]	ANOVA	Analysis of repeated measures in antidepressant effects over time	Mental Health
[123]	ANOVA	Adjusting for random baseline imbalances in clinical trials	Clinical Trials
[126]	ANCOVA	Handle missing data and apply covariate adjustment in weight-loss trials	Weight-Loss Trials
[52]	ANCOVA	Incorporates social network structure and opinion for election forecasting	Election Prediction

and independence between random and fixed effects. While diagnostic checks, transformations, and robust estimation techniques can address some violations, these assumptions inherently restrict their use in datasets with non-linear relationships or non-normal distributions. Computational challenges with large datasets or complex random-effects structures can be mitigated using efficient algorithms like Expectation-Maximization or parallel computing. However, LMMs struggle with data missing not at random (MNAR), as the mechanism depends on unobserved variables, and biased estimates may result when the independence of random and fixed effects is violated. Additionally, interpreting random effects in complex hierarchical models can be challenging, limiting subject-specific inferences. These limitations underscore the need to carefully evaluate assumptions and choose appropriate methods for specific research contexts.

2.2.2 Generalized Estimating Equations

Generalized Estimating Equations (GEEs), introduced by Liang and Zeger in 1986, are an extension of generalized linear models (GLMs) designed to handle correlated or clustered data, such as repeated measures or longitudinal observations. Unlike LMMs, which explicitly include random effects to account for individual-level variability, GEEs focus on estimating population-averaged effects, making them ideal for studies where the primary interest lies in overall population trends rather than subject-specific inferences. The general form of a GEE is: $g(\mu_{ij}) = X_{ij}\beta$, where $g(\cdot)$ is the link function (e.g., logit for binary outcomes, log for count data, identity for continuous data), μ_{ij} represents the mean response for the j-th observation of the i-th cluster, X_{ij} is the covariate matrix, and β denotes the regression coefficients. GEEs account for within-cluster correlation by specifying a working correlation structure, such as exchangeable, autoregressive, or unstructured. While the correct specification of this structure enhances efficiency, GEEs remain robust even if it is misspecified, providing consistent estimates of regression coefficients.

GEEs are highly flexible, accommodating various outcome types—binary, count, and continuous—via appropriate link functions. They are computationally efficient as they avoid explicit random-effects modeling and are robust to correlation structure misspecification, ensuring reliable population-level inferences. However, they assume data are Missing Completely at Random (MCAR), a stricter condition than the MAR assumption used in LMMs. GEEs focus on population-averaged effects, limiting their ability to provide individual-level inferences or subject-specific predictions. Additionally, their efficiency relies on correctly specifying the working correlation matrix, and they cannot nat-

Table 3. Applications of LMMs in Clinical Trials.

Study	Method	Application	Therapeutic Area
[81]	LMM	Model patient responses over time, introducing random effects for patient-specific trajectories	Psychiatry
[85]	GEE	Analyzing repeated measurements of respiratory outcomes, establishing a foundation for their use in public health research	Respiratory Outcomes
[161]	GEE	Binary data analysis, focusing on treatment adherence in clinical trials	Clinical Trials
[97]	GEE	Modeling correlated binary outcomes in healthcare infection rate studies	Infection Control
[145]	LMM	Monitoring HbA1c levels in diabetes trials	Diabetes
[37]	LMM	Phase II trial and pharmacokinetic evaluation of cytosine arabinoside for leptomeningeal metastases from breast cancer	Oncology
[144]	LMM	Estimate the reliability of repeated measurements in clinical trial data for schizophrenia treatment	Psychiatry
[56]	LMM	Drug trials evaluating the efficacy of antidepressants across multiple time points	Psychiatry
[43]	LMM	Modeling tumor size reduction over time, handling irregularly spaced follow-up data	Oncology
[43]	GEE	Evaluate weight loss interventions by modeling repeated weight measurements over time	Weight Management
[120]	GEE	Investigated concordance between urine drug screens and self-reported cocaine use over time and across genders	Addiction Medicine
[21]	LMM	Analyzing fMRI group data and study brain activation patterns in response to stimuli	Neuroscience
[36]	GEE	Analyzed self-efficacy in treatment frameworks among psychology and management scholars	Psychology
[127]	GEE	Conducted a randomized controlled trial of CBT-AD to improve adherence and reduce depression among HIV-positive Latinos	Behavioral Medicine
[25]	GEE	Estimated the efficacy of preexposure prophylaxis for HIV prevention based on drug concentration thresholds	Epidemiology
[153]	LMM	Analysis of non-adherence to treatment in a randomized controlled trial comparing citalopram and reboxetine in treating depression	Psychiatry
[148]	LMM	Measuring anthelmintic drug efficacy for parasitologists	Parasitology
[22]	GEE	Examined how housing status influenced drug use patterns among street-involved youth in Canada	Public Health
[129]	LMM	Modeled QTc prolongation to evaluate the safety profile of olmesartan medoxomil	Cardiovascular Pharmacology
[58]	GEE	Performed a systematic review of drug efficacy studies for soil-transmitted helminthiases and advocated for individual data-sharing	Parasitology
[72]	LMM	Statistical analysis of intestinal lesion scores in studies of anti-coccidial drugs in chickens	Veterinary Parasitology
[57]	LMM	Design and analysis of mouse clinical trials for oncology drug development	Oncology
[74]	LMM	Systematic review and meta-analysis of in vivo efficacy of anti-malarial drugs against clinical Plasmodium vivax malaria in Ethiopia	Infectious Diseases
[68]	GEE	Evaluated the effects of skin-to-skin contact on newborn sucking and breastfeeding abilities in a quasi-experimental study	Neonatal Health
[89]	LMM	Assessed the efficacy of a digital integrative medicine intervention for cancer patients undergoing treatment	Oncology

urally handle complex random-effect structures, restricting their use in hierarchical settings. Despite these limitations, GEEs are widely applied in clinical and epidemiological studies for analyzing correlated data and modeling average treatment effects.

2.2.3 Modern Extensions of LMMs and GEEs

Linear Mixed-Effects Models (LMMs) and Generalized Estimating Equations (GEEs) are widely used for analyzing longitudinal and hierarchical data. However, complex relationships often require extensions such as Nonlinear Mixed-

Effects Models (NLMMs) and Generalized Additive Mixed Models (GAMMs). NLMMs extend LMMs by incorporating nonlinear functions to model complex biological processes like drug absorption and elimination in pharmacokinetics. These models take the form $Y_{ij} = f(X_{ij}, \beta, b_i) + \epsilon_{ij}$, where $f(\cdot)$ is a nonlinear function, β represents fixed effects, b_i accounts for random effects, and ϵ_{ij} captures residual errors. Tools such as NONMEM, Monolix, and the R package `nlme` facilitate the application of NLMMs, although their computational demands are significant [104, 27]. GAMMs combine smooth, nonparametric functions with mixed effects to model unknown or varying relationships. The general form of a GAMM is: $Y_{ij} = X_{ij}\beta + Z_{ij}b_i + \sum_k f_k(X_{ijk}) + \epsilon_{ij}$, where $\sum_k f_k(X_{ijk})$ represents smooth functions of predictors, and $Z_{ij}b_i$ captures the random effects. GAMMs are commonly used in environmental and disease progression studies but may face overfitting and computational challenges in large datasets [154].

NLMMs and GAMMs expand the flexibility of LMMs and GEEs by accommodating nonlinear and nonparametric relationships, making them invaluable for modeling complex processes in clinical and epidemiological research. However, careful consideration is needed to address computational demands and ensure model interpretability.

2.2.4 Modeling with Survival Outcomes

Accelerated Failure Time (AFT) models are commonly used to analyze time-to-event data by directly modeling the logarithm of survival times as a linear function of covariates. The general form of an AFT model is: $\log(T_i) = X_i\beta + \epsilon_i$, where T_i is the survival time for the i-th individual, X_i is the covariate vector, β is the vector of regression coefficients and ϵ_i is the error term. AFT models assume that covariates accelerate or decelerate the time to the event by a constant factor. For example, if $\beta > 0$, the covariate increases survival time by a factor e^{β}; if $\beta < 0$, it decreases survival time by the same factor. Common distributions for ϵ yield specific AFT models, such as the Weibull AFT or log-logistic AFT.

When longitudinal data is integrated with survival response, Joint Models (JMs) are used to address the dependency between longitudinal trajectories and time-to-event processes, offering a unified framework for analyzing such data. The general joint modeling framework consists of two submodels: the longitudinal submodel, given by $Y_i(t) = X_i^{\top}(t)\beta + Z_i^{\top}(t)b_i + \epsilon_i(t)$, where $Y_i(t)$ is the longitudinal outcome for subject i at time t, $X_i(t)$ and $Z_i(t)$ are design matrices for fixed (β) and random (b_i) effects, respectively, $\epsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$ represents residual errors, and the survival submodel $h_i(t|M_i(t), w_i) = h_0(t)\exp(w_i^{\top}\gamma + \alpha M_i(t))$, where $h_i(t)$ is the hazard function for subject i at time t, $h_0(t)$ is the baseline hazard, w_i is a vector of baseline covariates with coefficients γ, $M_i(t)$ is a function of the subject-specific longitudinal trajectory, linking the two submodels via the association parameter α. JMs provide more accurate estimates of survival outcomes and dynamic predictions of survival probabilities by incorporating longitudinal biomarker information, thus reducing bias compared to separate analyses. However, they rely on strong parametric assumptions about the underlying processes, which, if misspecified, can lead to biased results. Fitting JMs is computationally intensive, requiring specialized algorithms like the Expectation-Maximization (EM) algorithm or Bayesian Markov Chain Monte Carlo (MCMC) methods. Recent advancements have addressed these challenges by developing efficient algorithms, such as the R package JSM [156], which provides a semiparametric joint modeling framework and offers a maximum likelihood approach to fit these models.

While AFT models and the proportional hazards framework remain the most widely used for survival data, several extensions have been proposed to address more complex event structures. When the proportional hazards assumption is violated, alternatives such as time-varying coefficient models, stratified Cox models, Random Survival Forest and Gradient Boosting type methods provide flexible inference by allowing hazard ratios to change over time [62, 137, 50]. In studies with complex censoring patterns, including interval-censored, left-truncated, or competing-risk

Table 4. Applications of AFT and Joint Models in Clinical Trials.

Study	Method	Application	Therapeutic Area
[39]	AFT	Analyze long-term survivors in clinical trials, emphasizing applicability in data with short- and long-term survival times	Clinical Trials
[149]	AFT	Use of Weibull AFT models to evaluate treatment effects and compare survival times in cancer patients receiving different chemotherapy regimens	Oncology
[14]	AFT	Comparison of AFT models with proportional hazards models in censored data survival analysis in cancer trials	Oncology
[140]	JM	HIV studies to explore relationships between viral load and time to treatment failure	HIV Studies
[108]	JM	Alzheimer's disease trials to model the association between cognitive decline and time to dementia onset	Alzheimer's Disease
[114]	JM	Predict survival probabilities for cancer patients based on tumor progression biomarkers	Oncology

data, specialized likelihood-based and nonparametric estimators have been developed, such as the Fine–Gray subdistribution model for competing risks [41]. Furthermore, when participants may experience multiple or repeated events, recurrent-event models such as the Andersen–Gill counting-process formulation, Prentice–Williams–Peterson (PWP) total- or gap-time models, and Wei–Lin–Weissfeld (WLW) marginal models offer distinct strategies for capturing dependence among event recurrences [100]. Together, these approaches expand survival modeling beyond proportional hazards assumptions and accommodate the complexities frequently encountered in modern clinical trials.

2.3 Handling Missing data

Parametric methods such as ANOVA, ANCOVA, and mixed-effects models rely on explicit distributional assumptions and are particularly sensitive to how missing data arise. In most applications, data are assumed to be missing at random (MAR), that is, the probability of missingness depends only on observed quantities. Under this assumption, likelihood-based estimation or restricted maximum likelihood provides valid inference while using all available data [86, 94]. When data are missing not at random (MNAR), for instance, when dropout depends on unobserved outcomes, bias can occur even with correct model specification. In such settings, models that explicitly link the missingness process to the outcome, such as selection, pattern-mixture, or shared-parameter formulations, are required [26]. Because MNAR assumptions cannot be empirically verified, sensitivity analyses are essential for assessing robustness. Delta-adjustment, tipping-point, and Bayesian prior–based approaches provide interpretable ways to examine how treatment estimates change under alternative assumptions. Clear documentation of the assumed mechanism and sensitivity framework is therefore critical for credible parametric inference.

3. NONPARAMETRIC TESTS

Nonparametric tests have become essential tools in clinical trials, particularly when the data deviate from the assumptions required by parametric methods, such as normality or homogeneity of variances. Nonparametric frameworks are particularly useful when the aim is to compare treatment distributions robustly without relying on parametric assumptions, emphasizing inference on ranks or medians rather than means. These methods are robust and versatile, making them well-suited for analyzing ordinal, skewed, or small-sample data, as well as datasets with outliers. In clinical trials, nonparametric tests are often used to evaluate treatment efficacy in scenarios where the primary interest lies in comparing distributions or ranks rather than means. For example, in trials with repeated measures or longitudinal outcomes, extensions of classical nonparametric tests, such as the Friedman test [45], Brunner-Munzel test [15], and nparLD [98], provide powerful alternatives to parametric counterparts. Additionally, in studies where only changes from baseline to the final endpoint are available, rank-based methods such as the Wilcoxon signed-rank test or Mann-Whitney U test are frequently applied. These methods are particularly advantageous in early-phase or exploratory trials with limited sample sizes, offering flexibility in handling non-normal or ordinal outcomes. However, despite their robustness, nonparametric tests may exhibit lower power than parametric methods when parametric assumptions are satisfied, and their interpretation can be less straightforward, particularly in complex longitudinal settings. By addressing these challenges and providing distribution-free inference, nonparametric methods play a vital role in efficacy testing, especially in studies involving diverse patient populations or unconventional outcome measures.

3.1 Cross-Sectional Nonparametric Tests

While parametric methods such as t-tests and ANOVA are widely used for efficacy testing, they rely on assumptions of normality and homogeneity of variance that may not hold in real-world clinical data. When these assumptions are violated – such as in the presence of skewed distributions, ordinal outcomes, or small sample sizes – nonparametric methods provide a robust alternative. The following section explores cross-sectional nonparametric tests, including the Wilcoxon Signed-Rank Test and the Mann-Whitney U Test, highlighting their advantages and applications in clinical trial settings.

3.1.1 Tests for Comparing Two Groups

The Wilcoxon Signed-Rank Test [152] is a non-parametric alternative to the paired t-test, used to compare paired observations when parametric assumptions are violated. It tests whether the median of the differences between paired values is zero by ranking the absolute values of the differences and summing the signed ranks. The test statistic is calculated as $W = \sum_{i=1}^{n} R_i \cdot \text{sgn}(D_i)$, where $D_i = X_i - Y_i$ and $\text{sgn}(D_i)$ is the sign of the difference. For small samples, critical values are from exact tables, while for large samples, the statistic approximates a normal distribution. This test is robust for small sample sizes and ordinal data but assumes symmetry in the differences, which if violated, can lead to biased results. It is less powerful than the paired t-test for normally distributed data.

The Mann-Whitney U Test [88] compares two independent groups, assessing if one group tends to have larger values than the other. It is particularly useful for ordinal data or skewed distributions. The U statistic is calculated based on ranks of the pooled data, with the formula: $U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum_{i=1}^{n_1} R(X_i)$, where $R(X_i)$ is the rank of X_i in the combined sample. The test assumes that the groups are independent and that the distributions have the same shape and scale. For small sample sizes, critical values are obtained from exact tables, while larger samples

use a normal approximation. It is less efficient than the t-test for normally distributed data, trading power for robustness. Furthermore, it assumes that the groups are independent and the distributions of the two groups have the same shape and scale; otherwise, the results may reflect differences in spread rather than central tendency.

Nonparametric tests for efficacy in a two-group setup has been widely studied in the literature, including [5], who evaluated blood pressure changes in antihypertensive drug trials, while [53] compared pain scores and skin improvement in crossover and dermatology trials, respectively. Other studies assessed the accuracy of diagnostic tools [82], cholesterol level changes in nutritional research [65], genetic and biomarker disclosure process [92] and progression-free survival in oncology [53].

3.1.2 Tests for Comparing More Than Two Groups

The Kruskal-Wallis (KW) test, developed by Kruskal and Wallis in 1952, is a non-parametric alternative to one-way Analysis of Variance (ANOVA). It is used to compare medians across two or more independent groups, particularly when the assumption of normality or homogeneity of variances is violated. Unlike ANOVA, the Kruskal-Wallis test ranks the data instead of analyzing raw values, making it robust against outliers and non-normal distributions. Consider k independent groups with sample sizes $n_1, n_2, \ldots, n_k$ and combined total observations $N = \sum_{i=1}^{k} n_i$. Let R_{ij} denote the rank of the j-th observation in group i, where all observations are ranked jointly across groups. The test statistic H is calculated as $H = 12(N(N+1))^{-1} \sum_{i=1}^{k} (T_i^2/n_i) - 3(N+1)$, where $T_i = \sum_{j=1}^{n_i} R_{ij}$ is the sum of ranks for the i-th group. Under the null hypothesis (H_0), which assumes that all groups are sampled from the same distribution, H approximately follows a chi-squared distribution with $k-1$ degrees of freedom for large N. When H_0 is rejected, it indicates significant differences among the group medians. While the test detects differences among groups, it does not identify which specific groups differ. Post-hoc pairwise comparisons must be conducted to pinpoint group differences, often using adjusted rank-based tests or Bonferroni corrections.

3.1.3 Tests for Association Between Categorical Variables

The Mantel-Haenszel (MH) test, introduced by Mantel and Haenszel in 1959, is a non-parametric statistical method used to evaluate associations between two categorical variables while controlling for a confounding variable. It is widely used in clinical trials and epidemiology to analyze stratified data, particularly in cases where the data are organized into contingency tables across strata. Given K strata, each represented by a 2×2 table, the MH test combines information across all strata to compute a pooled odds ratio and test for homogeneity or association. For each stratum k, the contingency table is represented as:

	Exposed (E)	Unexposed (U)
Case (C)	a_k	b_k
Control (Co)	c_k	d_k

The MH estimate of the common odds ratio ($\hat{\theta}_{MH}$), with $n_k = a_k + b_k + c_k + d_k$ as the total sample size for the k-th stratum, is calculated as $\hat{\theta}_{MH} = \sum_{k=1}^{K}(a_k d_k/n_k)/\sum_{k=1}^{K}(b_k c_k/n_k)$. The Mantel-Haenszel test statistic is computed as: $\chi^2_{MH} = \left(\sum_{k=1}^{K}(a_k - E_k)\right)^2/\sum_{k=1}^{K} V_k$, where E_k is the expected value of a_k under the null hypothesis of no association and V_k is its variance. The test statistic χ^2_{MH} follows a chi-squared distribution with 1 degree of freedom under the null hypothesis. A significant result suggests an association between the exposure and the outcome after controlling for the stratifying variable. However, the test assumes that the odds ratio is consistent across strata, which may not always hold in practice. If the homogeneity assumption is violated, the pooled estimate may be misleading. Furthermore, the test is limited to 2×2 contingency tables and cannot accommodate complex multilevel or continuous data without extensions. It also requires sufficient sample sizes within each stratum for reliable results.

3.1.4 Tests with Survival Data

The Log-Rank test, first introduced by Mantel in 1966, is a nonparametric statistical method used to compare survival distributions between two or more groups. The test is based on the null hypothesis (H_0) that there is no difference in the survival distributions between groups. Consider k groups, and let $n_i(t)$ and $d_i(t)$ denote the number of subjects at risk and the number of events observed at time t in group i, respectively. The observed (O_i) and expected (E_i) event counts for each group are calculated as follows: $E_i(t) = (n_i(t)/n(t)) \cdot d(t)$, where $n(t) = \sum_{i=1}^{k} n_i(t)$ is the total number of subjects at risk at time t, and $d(t) = \sum_{i=1}^{k} d_i(t)$ is the total number of events at time t. The observed and expected values are then summed across all time points. The test statistic is given by $\chi^2 = \left(\sum_{i=1}^{k}(O_i - E_i)\right)^2/\sum_{i=1}^{k} V_i$, where V_i is the variance of O_i under the null hypothesis, calculated as $V_i = (\sum_t n_i(t) \cdot (n(t) - n_i(t)) \cdot d(t) \cdot (n(t) - d(t)))/(n(t)^2 \cdot (n(t) - 1))$. Under H_0, the test statistic χ^2 follows a chi-squared distribution with $k-1$ degrees of freedom. For $k = 2$, the test reduces to a single degree of freedom test comparing two groups. The Log-Rank is a robust, distribution-free method for comparing survival curves and it retains good power when the hazard ratios between groups are proportional. However, violations of the proportional hazards assumption can lead to biased results. The test is also sensitive to censoring patterns; unbalanced censoring across groups can distort the results. Moreover, it does not adjust for covariates, requiring the use of stratified methods for more complex analyses.

Gray's test, introduced in [55], is a non-parametric method for comparing cumulative incidence functions in the presence of competing risks. In clinical trials and epidemiological studies, competing risks occur when an individual

Table 5. Applications of the Kruskal-Wallis (KW) and Mantel-Haenszel (MH) Tests in Clinical Trials.

Study	Method	Application	Therapeutic Area
[46]	MH	Meta-analysis comparing ceftriaxone with β-lactams in febrile neutropenia using the Peto-modified MH method to assess relative efficacy	Antibiotic Therapy
[87]	KW	Compared the efficacy of various preventive drugs during the course of preventive migraine treatment.	Neurology
[138]	MH	Meta-analysis to assess the efficacy and safety of combining Vandetanib with chemotherapy in advanced non-small cell lung cancer patients	Oncology
[30]	MH	Compare clinical success rates of moxifloxacin against pooled active comparator treatments in secondary peritonitis	Antibiotic Therapy
[112]	MH	Multicenter trials with binary outcomes to summarize data obtained from Early External Cephalic Version (EECV) trials published in 2003 across different strata	Multicenter Trials
[134]	MH	Infertility treatments clinical trials, emphasizing its simplicity and suitability for crossover designs	Infertility Treatments
[19]	KW	Compare the inpatient length of stay (LOS) between different patient samples, exploring the relationship between LOS, treatment benefit, and adverse events.	Healthcare, Inpatient Care
[124]	MH	Post-hoc analysis evaluating the efficacy of Lasmiditan for treating migraines in patients with cardiovascular risk factors	Neurology
[107]	KW	Explored factors affecting patient participation in clinical research to assess differences in willingness across groups.	Clinical Research
[13]	KW	Evaluated the efficacy of hydroxychloroquine in preventing illness compatible with Covid-19 or confirmed infection when used as postexposure prophylaxis.	Infectious Diseases
[67]	MH	Systematic review and meta-analysis on integrated care's impact on outcomes after acute coronary syndrome	Cardiovascular Research

is at risk of experiencing more than one mutually exclusive event, such as death from different causes. Unlike the Log-Rank test, which assumes all events are of the same type, Gray's test accounts for the sub-distribution of competing risks. The cumulative incidence function (CIF), denoted as $F_j(t)$, represents the probability of experiencing event j by time t, considering the presence of other competing events. Gray's test assesses whether the CIFs differ significantly between groups for a specific event type j. The test statistic for Gray's test is derived from weighted differences in observed and expected event counts for each group. For k groups, let $d_{ij}(t)$ denote the number of events of type j at time t in group i, and $n_i(t)$ denote the number of individuals at risk in group i at time t. The weighted observed-minus-expected difference is computed as $S(t) = \sum_{i=1}^{k} w(t) \cdot [d_{ij}(t) - E_{ij}(t)]$, where $E_{ij}(t)$ is the expected number of events in group i under the null hypothesis, and $w(t)$ is a weight function. The variance of $S(t)$ is estimated to compute the test statistic $\chi^2 = S(t)^2 / \mathrm{Var}[S(t)]$. Under the null hypothesis, χ^2 follows a chi-squared distribution with $k-1$ degrees of freedom. This test is found to be particularly valuable in clinical trials where competing risks are prominent, such as cancer studies with multiple causes of mortality. However, Gray's test is sensitive to the choice of weights, and assumes independence between competing events and censoring, which may not always hold in practice. Additionally, the test does not adjust for covariates, requiring extensions such as Fine and Gray's regression model for more complex analyses.

3.2 Longitudinal Nonparametric Tests

While cross-sectional nonparametric tests provide robust alternatives to parametric methods for single-time-point comparisons, many clinical trials involve repeated measurements over time. In such cases, traditional rank-based methods may fail to account for within-subject correlations, requiring specialized nonparametric techniques for longitudinal data. These methods, such as the Friedman Test, Brunner-Munzel Test, the nparLD framework and the Longitudinal Rank Sum Test, extend nonparametric inference to repeated measures settings, enabling more flexible and assumption-free analysis of treatment effects over time. The following section explores these approaches, their applications, and their advantages in handling non-normal, ordinal, and irregularly spaced longitudinal data.

3.2.1 Friedman Test

The *Friedman Test*, introduced by [45], is a nonparametric method designed for analyzing repeated measures or matched group data across multiple treatments. The

test ranks observations within each subject and evaluates differences in ranks across treatments, making it robust to non-normal distributions. For n subjects and k treatments, let R_{ij} denote the rank of the j-th treatment for the i-th subject. The test statistic is computed as: $Q = \frac{12n}{k(k+1)} \sum_{j=1}^{k} \left(\bar{R}_j - \frac{k+1}{2}\right)^2$, where $\bar{R}_j = \frac{1}{n} \sum_{i=1}^{n} R_{ij}$ is the average rank of treatment j. Under the null hypothesis that all treatments have identical distributions, Q approximately follows a chi-squared distribution with $k-1$ degrees of freedom when n is large. A significant result suggests differences in treatment effects. The Friedman Test is particularly suitable for ordinal data or non-normally distributed outcomes, where traditional parametric methods like repeated measures ANOVA are inappropriate.

The Friedman Test offers simplicity and robustness as its primary advantages. It does not assume normality and is straightforward to implement, making it accessible for small datasets and early-stage studies. However, it comes with notable limitations. The test assumes balanced data, which means every subject must have observations across all time points or treatments—a condition rarely met in real-world longitudinal studies with missing data. Additionally, it treats measurements as independent ranks within subjects, ignoring temporal trends or within-subject variability, which are critical in longitudinal studies. Despite these limitations, the Friedman Test remains a foundational method for small-scale repeated measures experiments, especially in fields like nutrition and behavioral studies.

3.2.2 Brunner-Munzel Test

The *Brunner-Munzel Test*, introduced by [15], is a rank-based nonparametric method designed to compare two groups while allowing for unequal variances and non-normal distributions. Unlike the Wilcoxon Rank-Sum Test, the Brunner-Munzel Test does not assume homogeneity of variances, making it robust in situations where group variances differ. For independent samples $X_1, \ldots, X_n$ from group 1 and $Y_1, \ldots, Y_m$ from group 2, the test statistic evaluates the probability $P(X < Y) + 0.5\,P(X = Y)$, interpreted as the stochastic dominance of one group over the other. The test statistic T is calculated as: $T = \frac{\hat{\Delta}}{\sqrt{\hat{\sigma}^2}}$, where $\hat{\Delta}$ is the difference in rank-based means between the two groups, and $\hat{\sigma}^2$ estimates the variance of the rank means. Under the null hypothesis of no difference between the groups, T follows a standard normal distribution asymptotically. The Brunner-Munzel Test is particularly useful in comparing treatment effects where variability differs substantially between groups, such as in clinical trials with heterogeneous populations.

The Brunner-Munzel Test offers significant advantages over traditional rank-based methods. Its ability to handle unequal variances makes it particularly robust in practical scenarios, such as comparing treatment efficacy in diverse patient populations or analyzing data with extreme outliers. Additionally, it retains the benefits of nonparametric methods, including robustness to non-normality and suitability for ordinal outcomes. However, the test is limited to two-group comparisons, making it less versatile for studies involving more than two treatments or time points. Moreover, like many rank-based tests, it does not explicitly incorporate temporal dependencies or repeated measures, which restricts its applicability in longitudinal settings. Despite these limitations, the Brunner-Munzel Test remains a valuable tool in scenarios requiring robust, nonparametric two-sample comparisons.

3.2.3 nparLD Framework

The *nparLD framework*, introduced by [98] is a nonparametric rank-based approach specifically designed for the analysis of longitudinal and repeated measures data. It extends traditional rank-based methods by using pseudo-rank transformations, which preserve the ordinal structure of the data while accounting for repeated measures. For a dataset with n subjects, t time points, and k treatment groups, let Y_{ij} represent the observation for subject i at time j. The pseudo-rank R^*_{ij} of Y_{ij} is calculated based on its rank relative to all observations across time points and groups. Using these pseudo-ranks, the framework evaluates treatment, time, and interaction effects through F-type test statistics. The F-type statistic for a given effect is expressed as: F = Between-group pseudo-rank variability/Residual pseudo-rank variability, where the numerator captures the variability explained by the effect (e.g., group, time, or interaction), and the denominator accounts for residual variability. For example, the test for a group effect compares the average pseudo-ranks across treatment groups, while controlling for time and subject variability. The null hypothesis of no effect is tested using permutation or large-sample asymptotic methods, with the F-statistic approximated by an F-distribution under the null.

The nparLD framework offers several advantages. Its ability to handle unbalanced data and missing observations makes it particularly suitable for real-world longitudinal studies, where dropout and irregular follow-up times are common. Additionally, it accommodates both ordinal and continuous outcomes, providing robust results even in the presence of non-normal data or outliers. The inclusion of interaction effects, such as group-by-time interactions, further enhances its utility for complex study designs. However, there are notable limitations. The need for multiplicity adjustments when testing multiple hypotheses can reduce statistical power, and the computational demands of pseudo-rank transformations and F-statistic calculations increase with larger datasets. While nparLD is robust for moderate-to-large sample sizes, its power may be limited in small-sample settings. Despite these challenges, nparLD has been widely adopted in clinical trials, with applications ranging from the evaluation of repeated measures of biomarkers to studies of vaccine efficacy and weight-loss interventions.

3.2.4 Longitudinal Rank-Sum Test (LRST)

The *Longitudinal Rank-Sum Test (LRST)*, introduced by [158], is a nonparametric method developed to evaluate treatment effects across multiple longitudinal endpoints in clinical trials. Unlike traditional rank-based methods, the LRST accounts for the complexity of repeated measures and multi-endpoint designs while maintaining robustness against non-normal distributions and outliers. Let x_{itk} and y_{jtk} represent the changes from baseline for subjects in the control and treatment groups, respectively, where i and j index subjects, t denotes time, and k represents outcomes. Observations are ranked across all subjects, time points, and outcomes, with ranks R_{itk} and R_{jtk}. The test statistic is calculated as $T_{\text{LRST}} = (\bar{R}_{y\cdots} - \bar{R}_{x\cdots})/\sqrt{\widehat{\text{Var}}(\bar{R}_{y\cdots} - \bar{R}_{x\cdots})}$, where $\bar{R}_{y\cdots}$ and $\bar{R}_{x\cdots}$ are the mean ranks for the treatment and control groups, respectively, aggregated across all time points and outcomes. Under the null hypothesis of no treatment effect, T_{LRST} asymptotically follows a standard normal distribution. LRST has also been developed for multi-arm clinical trials by [49, 51].

The LRST offers several advantages, making it a valuable tool in modern clinical trials. Its rank-based approach is robust to outliers and non-normal data distributions, which are common in real-world datasets. By simultaneously evaluating multiple longitudinal endpoints, the LRST reduces the need for multiplicity adjustments, enhancing statistical power while maintaining control of Type I error rates. Additionally, it handles missing data and irregular follow-ups efficiently, ensuring flexibility in complex trial designs. However, the test relies on large-sample approximations for its validity, and its performance in small-sample studies may require further evaluation. Despite these limitations, the LRST has shown strong applicability in neurodegenerative disease and oncology trials, where multiple outcomes such as motor function and cognitive performance are monitored over time, providing a comprehensive evaluation of treatment efficacy.

3.3 Handling Missing Data

Nonparametric and rank-based methods are sensitive to missing data because their validity depends on the complete ordering of observations. When data are missing completely at random (MCAR), complete-case analysis remains unbiased but less efficient. Under missing at random (MAR) assumptions, validity can be maintained through several approaches. Inverse Probability Weighting (IPW) reweights observed cases by the inverse of their estimated response probabilities, while Multiple Imputation (MI) replaces missing values with plausible draws from predictive models and combines results across imputations to incorporate uncertainty [118]. Extensions such as augmented IPW and weighted-rank procedures improve small-sample efficiency and robustness [3, 70]. When data are missing not at random (MNAR), rank-based inference alone cannot guarantee unbiasedness; combining these approaches with targeted sensitivity analyses remains essential for assessing robustness [84].

Table 6. Applications of Nonparametric Methods in Clinical Trials (Sorted by Year).

Study	Method	Application	Therapeutic Area
[7]	Friedman	Panic and agoraphobia scale in a clinical trial	Psychiatry
[75]	nparLD	Resistance to Phytophthora crown rot in cucumbers	Agriculture
[116]	Brunner-Munzel	Lesion-symptom mapping in cognitive neuroscience	Neuroscience
[28]	Friedman	Immune cell response to pemetrexed in pancreatic adenocarcinoma	Oncology
[64]	Brunner-Munzel	Ketamine use in refractory status epilepticus	Neurology
[141]	Brunner-Munzel	Degradation mechanisms in autophagy systems	Cellular Biology
[80]	Friedman	Stress-strength estimation in clinical trials	General Medicine
[76]	nparLD	Automatic detection of major depressive disorder using electrodermal activity	Mental Health
[83]	nparLD	Linkage between the I-3 gene for Fusarium wilt resistance and bacterial spot sensitivity in tomato	Agriculture
[142]	Friedman	Compare the efficacy of 19 anticancer compounds on HNSCC cell lines	Oncology
[6]	Friedman	Edetate disodium-based chelation for critical limb ischemia	Diabetes
[2]	nparLD	Exergame training for older adults	Geriatrics
[136]	nparLD	Metabolic adaptations to targeted therapies in uveal melanoma	Oncology
[121]	nparLD	Impact of isolation on mental health of athletes during COVID-19	Mental Health
[128]	Brunner-Munzel	Sporozoite vaccine efficacy for malaria prevention	Infectious Diseases
[90]	Friedman	Ethanolamine oleate injection for postoperative pain	Oral Surgery
[119]	Brunner-Munzel	Depression and anxiety in ischemic stroke patients	Neurology
[159]	nparLD	Biomarkers in gingival crevicular fluid during menopause	Periodontology
[4]	Friedman	Comparison of pain block methods in bariatric surgery	Anesthesiology
[93]	Friedman	Epidural prolotherapy versus steroids for chronic pain	Pain Management

4. BAYESIAN MODELS

Bayesian methods have become indispensable tools for analyzing longitudinal data in clinical trials, offering a probabilistic framework that seamlessly integrates prior knowledge with observed data. By explicitly modeling uncertainty, Bayesian approaches provide a comprehensive understanding of treatment effects, making them particularly well-suited for efficacy testing in complex and high-stakes settings. Bayesian approaches serve both confirmatory and decision-analytic objectives, integrating prior evidence with observed data to produce interpretable posterior estimates for efficacy or safety outcomes. These methods excel in handling challenges such as small sample sizes, hierarchical data structures, and irregularly sampled longitudinal data, all of which are common in clinical trials. Despite their numerous advantages, Bayesian methods face challenges, particularly in terms of computational demands and the selection of appropriate priors. However, advancements in scalable inference algorithms, such as variational inference and MCMC methods, continue to mitigate these issues, broadening the scope of Bayesian methods in clinical trials.

4.1 Bayesian Hierarchical Mixed Model

Bayesian Hierarchical Models (BHMs) are foundational tools in the analysis of longitudinal clinical trial data. These models account for multi-level data structures, such as repeated measurements nested within subjects, by introducing random effects at different levels of the hierarchy. The Bayesian framework incorporates prior distributions over model parameters, enabling the explicit representation of uncertainty. A typical BHM for longitudinal data can be expressed as $y_{ij} = \beta_0 + \beta_1 t_{ij} + b_i + \epsilon_{ij}$, where y_{ij} is the observed outcome for subject i at time j, β_0 and β_1 are fixed effects representing the population-level intercept and slope, $b_i \sim \mathcal{N}(0, \sigma_b^2)$ represents the random effect for subject i, $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ is the residual error, and t_{ij} represents time. In the Bayesian framework, prior distributions are specified for all parameters. For example, $\beta_0, \beta_1 \sim \mathcal{N}(0, \tau^2)$, $\sigma_b^2, \sigma^2 \sim$ Inverse-Gamma(a, b). The inclusion of prior information allows the incorporation of external knowledge or historical data, further enhancing model robustness. Posterior distributions of the parameters are obtained using MCMC methods or variational inference, enabling uncertainty quantification for all components of the model. However, the reliance on MCMC methods often requires significant computational resources. Furthermore, careful selection of prior distributions is essential, as improper priors may introduce bias into the model estimates. Interpretation of results, while probabilistically robust, can also be challenging in highly hierarchical settings.

4.2 Bayesian Nonparametrics and Dynamic Bayesian Networks

Bayesian nonparametric models extend Bayesian methods to settings where the complexity of the model is not fixed a priori but can grow with the data. These models are particularly useful in clinical trials for longitudinal data, where the underlying distributions or clusters may not be well-defined. Bayesian nonparametrics rely on stochastic processes, such as the Dirichlet process (DP) and Gaussian processes (GP), to construct flexible models. A foundational Bayesian nonparametric model is the Dirichlet Process Mixture Model (DPMM), which allows clustering without prespecifying the number of clusters. The DPMM is represented as $G \sim \text{DP}(\alpha, G_0)$, where G is the random probability measure, α is the concentration parameter controlling cluster formation, and G_0 is the base distribution. Observations x_i are modeled as $x_i \sim F(\theta_i)$, $\theta_i \sim G$, where F is the likelihood function and θ_i are parameters drawn from the Dirichlet process. For longitudinal data, extensions like the Hierarchical Dirichlet Process (HDP) and Dependent Dirichlet Process (DDP) have been developed to model repeated measures and temporal dependencies. For example, the DDP allows the distribution G_t at time t to evolve over time $G_t \sim \text{DP}(\alpha, G_{t-1})$, capturing temporal relationships in longitudinal data. Bayesian Additive Regression Trees (BART) is another popular nonparametric model that partitions the data space using an ensemble of regression trees. BART offers a robust way to model complex, nonlinear relationships in longitudinal data. On the downside, these approaches are computationally intensive, requiring advanced sampling techniques. Furthermore, hyperparameters require careful tuning to ensure meaningful results.

Bayesian Gaussian Processes (GPs) infer distributions over functions, allowing them to adaptively capture the underlying structure in longitudinal trajectories. GPs are particularly well-suited for irregularly sampled data, a common feature in clinical trials. A GP models a set of observations $y = \{y_1, y_2, \ldots, y_n\}$ at corresponding times $t = \{t_1, t_2, \ldots, t_n\}$ as a realization from a multivariate normal distribution $y \sim \mathcal{N}(m(t), K(t, t'))$, where $m(t)$ is the mean function, often set to zero for simplicity, and $K(t, t')$ is the covariance function (or kernel) that encodes the similarity between observations at times t and t'. Common choices for $K(t, t')$ include the squared exponential kernel $K(t, t') = \sigma^2 \exp\big(-(t - t')^2/2\ell^2\big)$, where σ^2 is the variance and ℓ is the length scale, controlling the smoothness of the function. The Bayesian nature of GPs allows for uncertainty quantification in predictions. Given observed data (t, y), the posterior predictive distribution for a new time t^* is also Gaussian, that is, $p(y^*|t^*, t, y) \sim \mathcal{N}(\mu(t^*), \sigma^2(t^*))$, where $\mu(t^*) = K(t^*, t)K(t, t)^{-1}y$, $\sigma^2(t^*) = K(t^*, t^*) - K(t^*, t)K(t, t)^{-1}K(t, t^*)$. GPs handle irregularly sampled data seamlessly and can accommodate heteroscedastic noise by modifying the covariance structure. But, they face scalability challenges as the computational cost grows cubically with the number of data points due to the inversion of the covariance matrix. Sparse approximations, such as inducing points, have been developed to mitigate this issue but may compromise model accuracy. Additionally, hyperparameter

Table 7. Applications of Bayesian Nonparametric and Graphical Models in Clinical Trials.

Study	Method	Application	Therapeutic Area
[8]	GP	Model tumor growth dynamics in oncology trials, showing their adaptability to complex trajectories	Oncology
[151]	BHM	Phase I/II trials investigating the safety and efficacy of drug combinations, focusing on dose finding and exploring therapeutic activity	Dose Finding
[63]	DP	Nested Dirichlet process model to account for physician-patient interactions in cluster randomized trials, improving treatment effect assessment	Healthcare
[79]	BHM	Bayesian adaptive design for Phase I/II trials with delayed outcomes, jointly modeling efficacy and toxicity for dose escalation in oncology	Oncology
[60]	GP	Handle multivariate longitudinal data, enabling simultaneous modeling of multiple biomarkers in cardiovascular trials	Cardiovascular Research
[135]	BHM	Bayesian optimization for dose-finding to balance efficacy and toxicity in biologic agent trials	Oncology
[113]	DP	Regression discontinuity designs to assess treatment effects in clinical trials with thresholds	Clinical Trials Design
[102]	GP	Identify personalized optimal doses in Phase I/II clinical trials by modeling toxicity and efficacy based on patient biomarkers	Dose Finding
[33]	GP	Optimize nanoparticle formulations for drug delivery, improving encapsulation efficiency and therapeutic efficacy	Drug Delivery

tuning and kernel selection require careful consideration, as these can significantly impact model performance.

4.3 Handling Missing Data

A major strength of the Bayesian framework is its coherent treatment of missing data through probabilistic modeling. Missing values are treated as additional unknown parameters and are integrated over during posterior sampling, eliminating the need for ad hoc imputation [118]. When data are missing at random (MAR), data-augmentation or Gibbs-sampling algorithms naturally propagate uncertainty from incomplete observations into posterior estimates. For missing not at random (MNAR) scenarios, the Bayesian approach allows explicit modeling of the missingness mechanism using selection, pattern-mixture, or shared-parameter formulations [69, 26]. Sensitivity analyses can be incorporated directly into the Bayesian hierarchy by specifying alternative priors for missingness parameters, enabling robust inference under unverifiable assumptions. This unified treatment of uncertainty across both observed and unobserved data makes Bayesian methods especially attractive for longitudinal and adaptive clinical trial analyses.

5. DEEP LEARNING BASED METHODS

Deep learning has revolutionized data analysis across various domains, including clinical trials, by providing powerful tools to model complex and high-dimensional data. Unlike traditional statistical methods, deep learning models do not rely on predefined assumptions about data distribution or structure. Instead, they leverage neural networks to learn intricate patterns and relationships directly from the data, making them particularly suited for efficacy testing in longitudinal clinical trials with complex, unstructured, or high-dimensional datasets. Machine learning and deep learning approaches primarily target prediction and adaptive decision-making, using flexible, data-driven algorithms that capture nonlinearities and interactions without explicit distributional assumptions. Despite their flexibility and power, deep learning methods face challenges in clinical trial settings. These include high computational demands, a reliance on large datasets for training, and difficulties in interpreting model outputs. To address these limitations, recent advancements have focused on integrating deep learning with Bayesian frameworks, enabling uncertainty quantification and enhancing interpretability. Techniques like Bayesian Neural Ordinary Differential Equations (Neural ODEs) and Graph Neural Networks (GNNs) have further pushed the boundaries of deep learning applications in clinical research, enabling continuous-time modeling and the analysis of networked clinical data.

5.1 Deep Mixed Effects Models (DMEMs)

Deep Mixed Effects Models (DMEMs) integrate the hierarchical structure of traditional mixed-effects models with the flexibility of neural networks, enabling the modeling of complex, nonlinear relationships in longitudinal data. Mixed-effects models decompose the observed outcome for subject i at time j, y_{ij}, into fixed effects (population-level trends), random effects (subject-specific deviations), and residual errors. In DMEMs, this decomposition is enhanced with a neural network $f(x_{ij};\theta)$, which captures nonlinear relationships in high-dimensional covariates x_{ij}. The model is expressed as $y_{ij} = f(x_{ij};\theta) + b_i + \epsilon_{ij}$, where $b_i \sim \mathcal{N}(0, \sigma_b^2)$ represents the random effects specific to subject i, and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ denotes the residual errors. The neural network parameters θ are trained alongside random effect variance σ_b^2 and residual variance σ^2. The neural network $f(x_{ij};\theta)$

can take various forms, such as feedforward networks for static data, recurrent neural networks for sequential data, or convolutional networks for spatially structured data. The optimization of DMEMs often involves maximum likelihood estimation (MLE) or Bayesian approaches. DMEMs are particularly powerful for longitudinal data as they allow the estimation of subject-specific trajectories and provide robust predictions even in the presence of missing data or unbalanced study designs. However, DMEMs are computationally demanding, requiring substantial resources to optimize both the neural network and random effects parameters. The interpretability of DMEMs can also be challenging due to the black-box nature of neural networks, necessitating the use of feature attribution methods, such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations).Moreover, overfitting is a concern, particularly in smaller datasets, and regularization techniques must be carefully implemented.

5.2 Recurrent Neural Networks and Temporal Convolutional Networks

These belong to a class of neural networks designed to handle sequential and time-series data. They are particularly well-suited for longitudinal data in clinical trials because they allow for temporal dependencies between observations. Unlike traditional neural networks, recurrent neural networks (RNNs) include recurrent connections that create feedback loops, enabling the network to retain information from previous time steps in a hidden state. This structure makes RNNs powerful tools for modeling dynamic systems where the sequence of observations is crucial. Unlike RNNs, which process data sequentially, temporal convolutional networks (TCNs) leverage convolutional layers to capture temporal dependencies, making them computationally efficient and scalable for long sequences. TCNs are particularly suited for irregular sampling and complex time dependencies often occur.

The basic RNN model for a time-series sequence $x = \{x_1, x_2, \ldots, x_T\}$ computes the hidden state h_t at time t as $h_t = \sigma(W_h h_{t-1} + W_x x_t + b_h)$, where W_h and W_x are weight matrices for the hidden state and input, respectively, b_h is the bias vector, σ is the activation function, typically a hyperbolic tangent (tanh) or rectified linear unit (ReLU). The output y_t at time t is then computed as $y_t = \phi(W_y h_t + b_y)$, where W_y is the output weight matrix, b_y is the output bias, and ϕ is the output activation function. While RNNs can theoretically model long-term dependencies, in practice, they suffer from the vanishing and exploding gradient problem during training. To address these limitations, two specialized architectures have been developed: Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs). RNNs, especially LSTMs and GRUs, are powerful for handling sequential data with temporal dependencies. They excel in modeling complex longitudinal relationships and can naturally accommodate varying lengths of time series, making them suitable for clinical trials with irregular follow-ups. However, RNNs are computationally expensive, particularly for long sequences, due to the sequential nature of their computations. Training RNNs is challenging due to issues like vanishing gradients in vanilla RNNs and the need for extensive hyperparameter tuning. LSTMs and GRUs mitigate some of these challenges but at the cost of increased architectural complexity.

The fundamental architecture of TCNs includes the following features, namely, **Causal Convolutions**, which ensure that predictions at time t are only influenced by data from t and earlier, **Dilated Convolutions**, which introduce gaps between filter applications to expand the receptive field without increasing computational cost, **Residual Connections** which address vanishing gradient issues and improve gradient flow, and **Sequence Padding** to maintain the sequence length throughout the network, zero-padding is applied to the input. The output of a TCN is generated by applying successive convolutional layers, culminating in a prediction layer. This architecture makes TCNs highly parallelizable, unlike RNNs, which require sequential processing. Thus, TCNs do not suffer from the vanishing gradient problem inherent in RNNs, ensuring stable training. However, their reliance on fixed kernel sizes and dilation rates may require extensive hyperparameter tuning to achieve optimal performance. Additionally, both these approaches face an interpretability issue due to their black-box nature.

5.3 Graph Neural Networks (GNNs)

Graph Neural Networks (GNNs) are deep learning models designed to operate on graph-structured data, where nodes represent entities, and edges encode relationships between them. GNNs are increasingly applied to clinical trials, where data often involves complex relationships between patients, treatments, or time-series features. Unlike traditional deep learning models, which assume data independence, GNNs can learn from interdependent entities, making them well-suited for modeling networks of patients, molecular pathways, or clinical sites. For longitudinal clinical trial data, temporal extensions of GNNs, such as Temporal GNNs or Dynamic GNNs, incorporate time as an additional dimension to model evolving patient relationships or feature dynamics. However, GNNs are computationally expensive, especially for large graphs with dense connections. Training can be challenging due to issues like over-smoothing, where embeddings of all nodes become indistinguishable after several layers.

5.4 Federated and Reinforcement Learning with Deep Models

Federated learning (FL) is a decentralized approach to training deep learning models across multiple data sources without the need to centralize data. This is particularly valuable in clinical trials, where privacy concerns, regulatory constraints, and logistical challenges often prevent data from

Table 8. Applications of Deep Learning Methods in Clinical Trials. RL, Reinforcement Learning; TCN, Temporal Convolution Network; GNN, Graph Neural Network; RNN, Recurrent Neural Network; FL, Federated Learning; DMEM, Deep Mixed Effects Model.

Study	Method	Application	Therapeutic Area
[162, 163]	RL	Cancer clinical trial design	Oncology
[78]	RL	Sepsis management, recommending dynamic treatment adjustments based on evolving patient conditions	Critical Care
[59]	RNN	Use of multitask learning for mortality and length-of-stay prediction in critical care settings	Critical Care
[23]	TCN	Seizure detection using temporal graph CNNs	Neurology
[77]	TCN	Sepsis prediction using TCNs	Critical Care
[125]	FL	Develop machine learning models for brain tumor segmentation with multi-center imaging data, preserving patient privacy	Oncology
[29]	FL	Predicting COVID-19 outcomes using federated models	Infectious Diseases
[95]	RL	Potential applications of RL in ophthalmology	Ophthalmology
[18]	RNN	Estimation of Jadad's score for clinical trial robustness	Clinical Trial Methodology
[66]	DMEM	Mixture model for healthcare time-series data with Gaussian processes	Healthcare Analytics
[164]	RNN	Chronic kidney disease progression prediction using EHRs	Nephrology
[139]	RL	Online RL in oral health clinical trials	Oral Health
[157]	DMEM	Real-time monitoring in additive manufacturing with mixed-effects models	Manufacturing Processes
[101]	DMEM	Personalized prediction of Parkinson's progression using Gaussian processes	Neurology

being pooled across institutions or sites. In federated learning, models are trained locally on each participating site and periodically aggregated to form a global model. Extensions like personalized federated learning allow customization of global models to individual institutions, addressing heterogeneity across sites, without exposing sensitive patient data, ensuring compliance with privacy regulations like HIPAA and GDPR. However, communication overhead between institutions can be substantial, especially with frequent model updates. Data heterogeneity across sites can lead to suboptimal convergence or biased global models. Security concerns, such as model inversion attacks, must also be addressed to ensure participant confidentiality.

Reinforcement Learning (RL) is a framework for sequential decision-making, where an agent learns to maximize cumulative rewards through interactions with an environment. In the context of clinical trials, RL has been employed to optimize treatment strategies by modeling patient responses to interventions as a dynamic process. RL is particularly suited for longitudinal data analysis, where the effects of treatments unfold over time. RL offers the ability to personalize treatment strategies dynamically, adjusting decisions based on patient responses over time. However, RL requires substantial amounts of data to train effectively, which can be a challenge in clinical settings with limited patient samples.

5.5 Handling Missing Data

Missing data are pervasive in clinical trial settings, particularly in multimodal and longitudinal studies that integrate imaging, omics, and clinical outcomes. Deep learning models, while highly flexible, generally require complete input tensors, making missingness a key challenge for model reliability and inference. Traditional solutions rely on multiple imputation or inverse probability weighting (IPW) before model fitting, but recent approaches embed missingness handling directly within the network architecture. Generative models such as Variational Autoencoders (VAEs) and Generative Adversarial Imputation Networks (GAIN) reconstruct missing values by learning joint latent representations of observed and unobserved features, thus preserving biological structure and temporal coherence [160, 16]. For time-dependent data, Recurrent Neural Networks and Temporal Convolutional Networks have been extended to incorporate missing-indicator embeddings or time-gap encodings, enabling dynamic imputation during sequence learning [20, 17]. Despite these advances, most deep models assume data are missing at random (MAR); explicitly modeling missing not at random (MNAR) mechanisms remains an open challenge. Integrating deep generative frameworks with Bayesian or causal formulations provides a promising direction for handling informative dropout and achieving principled uncertainty quantification in high-dimensional clinical trial data.

6. CONCLUSION AND FUTURE DIRECTIONS

A conceptual integration framework summarizing the links between study design, endpoint type, and analytic approach is presented in Figure 1 to provide readers with a practical overview of the decision pathways discussed throughout this review. This schematic highlights how the

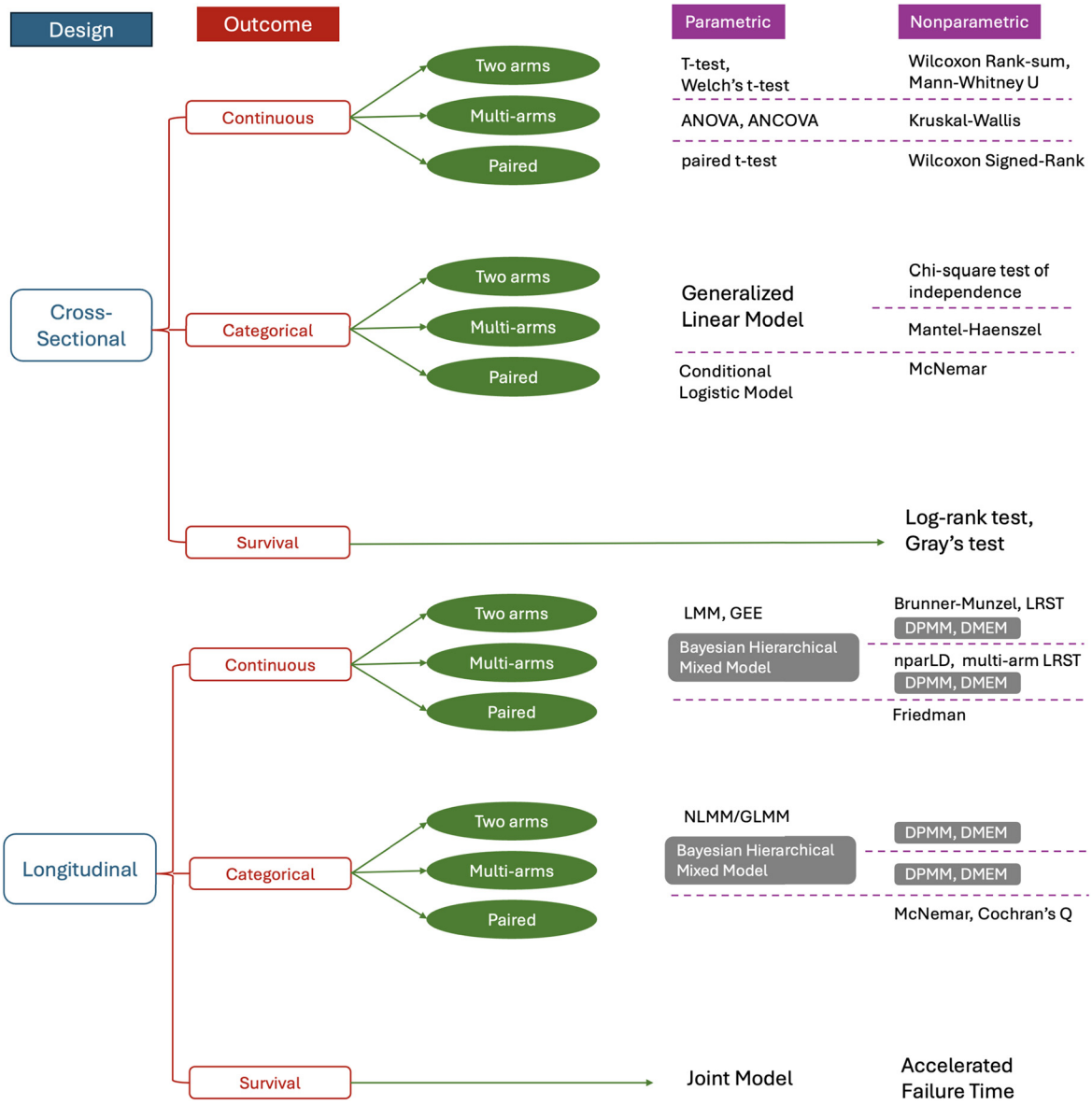

Figure 1: Schematic decision framework linking study design, outcome type, and analytic approach. The diagram guides readers from data structure (cross-sectional or longitudinal) through outcome characteristics (continuous, categorical, survival) to suitable parametric and nonparametric tests for two-arm, multi-arm, or paired comparisons.

choice of statistical test naturally follows from the structure of the data and the nature of the outcome, serving as a bridge between traditional inferential methods and emerging data complexities. Building on this framework, several challenges and opportunities remain that warrant further exploration. The growing complexity of clinical data, driven by the integration of multi-modal biomarkers, electronic health records, and real-world evidence, demands novel statistical and computational approaches that can handle high-dimensional, heterogeneous, and often incomplete datasets.

One key area of future work is the development of hybrid models that combine the strengths of parametric and nonparametric methods. While parametric methods offer efficiency and interpretability under ideal assumptions, nonparametric approaches provide robustness to deviations from these assumptions. Hybrid frameworks could leverage both perspectives, providing flexible yet interpretable solutions for efficacy testing. For example, the integration of rank-based methods with mixed-effects modeling could offer a promising avenue for analyzing complex longitudinal data.

Another promising direction is the advancement of Bayesian methodologies for efficacy testing. Although Bayesian methods have gained traction for their ability to in-

corporate prior information and quantify uncertainty, their application to high-dimensional and dynamic datasets is still evolving. The development of computationally efficient Bayesian frameworks, such as scalable MCMC algorithms or variational inference techniques, could enable their broader adoption in large-scale clinical trials.

In the realm of machine learning, deep learning models and reinforcement learning strategies have demonstrated potential in optimizing trial designs, predicting outcomes, and personalizing treatments. However, their utility in efficacy testing remains underexplored. Challenges such as lack of interpretability, risk of overfitting, and the need for large labeled datasets present significant barriers to their widespread application. Future research should focus on developing interpretable and domain-specific machine learning models tailored to clinical trial data, particularly those involving time-dependent outcomes. Moreover, the field lacks a consensus on the best practices for handling missing data in efficacy testing. While methods such as multiple imputation and mixed-effects modeling are widely used, their assumptions and limitations can vary significantly across trials. Further work is needed to develop robust, assumption-free approaches for handling missingness, particularly in longitudinal and adaptive trial designs.

Ensuring fairness in clinical efficacy testing has become increasingly critical as modern trials incorporate diverse patient populations and complex data sources. Statistical and algorithmic frameworks must therefore guard against biases that can arise from unequal subgroup representation, differential data quality, or model overfitting to majority cohorts. Recent strategies include stratified design and randomization, covariate adjustment for underrepresented groups, and reweighting or balancing techniques that equalize subgroup influence during estimation and prediction. In data-driven frameworks such as machine learning and deep learning, fairness-aware loss functions, adversarial debiasing, and post-hoc calibration methods have shown promise for improving equitable model performance across demographic groups. Integrating these fairness principles into efficacy analyses enhances the generalizability and ethical robustness of statistical inference, ensuring that emerging clinical evidence benefits all populations equitably.

Looking forward, several emerging research priorities are poised to shape the next generation of clinical efficacy methodology. One key direction involves developing robust and computationally efficient models for high-dimensional longitudinal data that integrate multimodal inputs such as imaging, genomics, and digital health records while preserving statistical validity [32, 145]. Another important challenge is balancing predictive accuracy with causal interpretability, as machine learning and deep learning frameworks become central to trial analytics [35, 12]. Integrative approaches that couple traditional inferential rigor with scalable representation learning, such as hybrid Bayesian–machine learning models, offer promising pathways toward interpretable yet flexible inference. The rapid rise of virtual and decentralized clinical trials (VCTs), enabled by telehealth, wearables, and remote monitoring, introduces new opportunities and challenges related to irregular data streams, device-based measurement error, and adherence variability [1, 143, 24, 34, 71]. Finally, ensuring fairness, transparency, and transportability in these data-rich settings is essential for generating equitable and generalizable evidence [133, 21]. Together, these directions underscore a shift toward methodological frameworks that are statistically rigorous, computationally adaptive, and ethically grounded, advancing the science of clinical evaluation in an era of data-driven and personalized medicine.

In summary, while the field has seen substantial progress, opportunities remain to address gaps in integrating advanced computational techniques, improving scalability and interpretability, and ensuring equitable and inclusive efficacy testing. Bridging these gaps will not only enhance the reliability of clinical trial outcomes but also pave the way for more personalized and effective healthcare solutions. By addressing these challenges, future research can ensure that statistical methodologies continue to evolve in step with the increasing complexity and scope of modern clinical trials.

Accepted 26 February 2026

REFERENCES

[1] Abadi, E., Segars, W. P., Tsui, B. M., Kinahan, P. E., Bottenus, N., Frangi, A. F., Maidment, A., Lo, J. and Samei, E. (2020). Virtual clinical trials in medical imaging: a review. *Journal of Medical Imaging* **7**(4) 042805.

[2] Adcock, M., Fankhauser, M., Post, J., Lutz, K., Zizlsperger, L., Luft, A. R., Guimarães, V., Schättin, A. and de Bruin, E. D. (2020). Effects of an in-home multicomponent exergame training on physical functions, cognition, and brain volume of older adults: a randomized controlled trial. *Frontiers in medicine* **6** 321.

[3] Akritas, M. G., Arnold, S. F. and Brunner, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *Journal of the American Statistical Association* **92**(437) 258–265. https://doi.org/10.2307/2291470. MR1436114

[4] Algyar, M. F. and Abdelsamee, K. S. (2024). Laparoscopic assisted versus ultrasound guided transversus abdominis plane block in laparoscopic bariatric surgery: a randomized controlled trial. *BMC anesthesiology* **24**(1) 133.

[5] Altman, D. G. (1991) *Practical Statistics for Medical Research*. Chapman and Hall/CRC.

[6] Arenas, I., Ujueta, F., Diaz, D., Yates, T., Olivieri, B., Beasley, R. and Lamas, G. (2019). Limb preservation using edetate disodium-based chelation in patients with diabetes and critical limb ischemia: an open-label pilot study. *Cureus* **11**(12).

[7] Bandelow, B., Brunner, E., Broocks, A., Beinroth, D., Hajak, G., Pralle, L. and Rüther, E. (1998). The use of the Panic and Agoraphobia Scale in a clinical trial. *Psychiatry Research* **77**(1) 43–49.

[8] Barber, D. and Williams, C. K. (2001). Gaussian Processes for Bayesian Modeling of Tumor Growth. *Neural Information Processing Systems*.

[9] Beasley, R., Harrison, T., Peterson, S., Gustafson, P., Hamblin, A., Bengtsson, T. and Fagerås, M. (2022). Evaluation of budesonide-formoterol for maintenance and reliever

therapy among patients with poorly controlled asthma: a systematic review and meta-analysis. *JAMA network open* **5**(3) 220615.

[10] Beaton, D. E. et al. (2002). Measuring the burden of musculoskeletal conditions. *Arthritis & Rheumatism*.

[11] Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American statistical Association* **82**(397) 112–122. MR0883340

[12] Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association* **112**(518) 859–877. https://doi.org/10.1080/01621459.2017.1285773. MR3671776

[13] Boulware, D. R., Pullen, M. F., Bangdiwala, A. S., Pastick, K. A., Lofgren, S. M., Okafor, E. C., Skipper, C. P., Nascene, A. A., Nicol, M. R., Abassi, M. et al. (2020). A randomized trial of hydroxychloroquine as postexposure prophylaxis for Covid-19. *New England journal of medicine* **383**(6) 517–525.

[14] Bradburn, M. J. et al. (2003). Survival analysis part II: Multivariate data analysis—an introduction to concepts and methods. *British Journal of Cancer*.

[15] Brunner, E. and Munzel, U. (2000). The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation. *Biometrics* **56**(4) 1173–1182. https://doi.org/10.1002/(SICI)1521-4036(200001)42:1<17::AID-BIMJ17>3.0.CO;2-U. MR1744561

[16] Camino, R. D., Hammerschmidt, C. A. and State, R. (2019). Improving missing data imputation with deep generative models. *arXiv preprint arXiv:1902.10666*.

[17] Cao, W., Wang, D., Li, J., Zhou, H., Li, L. and Li, Y. (2018). Brits: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems* **31**.

[18] Casy, T., Grasseau, A., Charras, A., Rouvière, B., Pers, J. -O., Foulquier, N. and Saraux, A. (2022). Assessing the robustness of clinical trials by estimating Jadad's score using artificial intelligence approaches. *Computers in Biology and Medicine* **148** 105851.

[19] Chazard, E., Ficheur, G., Beuscart, J. -B. and Preda, C. (2017). How to compare the length of stay of two samples of inpatients? A simulation study to compare type I and type II errors of 12 statistical tests. *Value in Health* **20**(7) 992–998.

[20] Che, Z., Purushotham, S., Cho, K., Sontag, D. and Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. *Scientific reports* **8**(1) 6085.

[21] Chen, G., Saad, Z. S., Britton, J. C., Pine, D. S. and Cox, R. W. (2013). Linear mixed-effects modeling approach to FMRI group analysis. *Neuroimage* **73** 176–190.

[22] Cheng, T., Wood, E., Nguyen, P., Kerr, T. and DeBeck, K. (2014). Increases and decreases in drug use attributed to housing status among street-involved youth in a Canadian setting. *Harm reduction journal* **11** 1–6.

[23] Covert, I. C., Krishnan, B., Najm, I., Zhan, J., Shore, M., Hixson, J. and Po, M. J. (2019). Temporal graph convolutional networks for automatic seizure detection. In *Machine learning for healthcare conference* 160–180. PMLR.

[24] Dahal, L., Ghojoghnejad, M., Vancoillie, L., Ghosh, D., Bhandari, Y., Kim, D., Ho, F. C., Tushar, F. I., Luo, S., Lafata, K. J. et al. (2025). XCAT 3.0: A comprehensive library of personalized digital twins derived from CT scans. *Medical Image Analysis* 103636.

[25] Dai, J. Y., Gilbert, P. B., Hughes, J. P. and Brown, E. R. (2013). Estimating the efficacy of preexposure prophylaxis for HIV prevention among participants with a threshold level of drug concentration. *American journal of epidemiology* **177**(3) 256–263.

[26] Daniels, M. and Hogan, J. (2008) *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. CRC Press. https://doi.org/10.1201/9781420011180. MR2459796

[27] Davidian, M. (2017) *Nonlinear models for repeated measurement data*. Routledge.

[28] Davis, M., Conlon, K., Bohac, G. C., Barcenas, J., Leslie, W., Watkins, L., Lamzabi, I., Deng, Y., Li, Y. and Plate, J. M. (2012). Effect of pemetrexed on innate immune killer cells and adaptive immune T cells in subjects with adenocarcinoma of the pancreas. *Journal of Immunotherapy* **35**(8) 629–640.

[29] Dayan, I., Roth, H. R., Zhong, A., Harouni, A., Gentili, A., Abidin, A. Z., Liu, A., Costa, A. B., Wood, B. J., Tsai, C. -S. et al. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. *Nature medicine* **27**(10) 1735–1743.

[30] De Waele, J. J., Tellado, J. M., Weiss, G., Alder, J., Kruesmann, F., Arvis, P., Hussain, T. and Solomkin, J. S. (2014). Efficacy and safety of moxifloxacin in hospitalized patients with secondary peritonitis: pooled analysis of four randomized phase III trials. *Surgical infections* **15**(5) 567–575.

[31] De Winter, J. C. F. (2013). Using the Student's t-test with extremely small sample sizes. *Practical Assessment, Research, and Evaluation* **18**(10) 1–12.

[32] Diggle, P. (2002) *Analysis of longitudinal data*. Oxford university press.

[33] Dong, S., Yu, H., Poupart, P. and Ho, E. A. (2024). Gaussian processes modeling for the prediction of polymeric nanoparticle formulation design to enhance encapsulation efficiency and therapeutic efficacy. *Drug Delivery and Translational Research* 1–17.

[34] Dorsey, E. R. and Topol, E. J. (2016). State of telehealth. *New England journal of medicine* **375**(2) 154–161.

[35] Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

[36] Elias, S. M., Barney, C. E. and Bishop, J. W. (2013). The treatment of self-efficacy among psychology and management scholars. *Journal of Applied Social Psychology* **43**(4) 811–822.

[37] Esteva, F. J., Soh, L. -T., Holmes, F. A., Plunkett, W., Meyers, C. A., Forman, A. D. and Hortobagyi, G. N. (2000). Phase II trial and pharmacokinetic evaluation of cytosine arabinoside for leptomeningeal metastases from breast cancer. *Cancer chemotherapy and pharmacology* **46** 382–386.

[38] Fagerland, M. W. et al. (2011). Statistical analysis of contingency tables. *BMC Medical Research Methodology* **11**(1) 47.

[39] Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*.

[40] Ferrario, C. M., Jessup, J., Chappell, M. C., Averill, D. B., Brosnihan, K. B., Tallant, E. A., Diz, D. I. and Gallagher, P. E. (2005). Effect of angiotensin-converting enzyme inhibition and angiotensin II receptor blockers on cardiac angiotensin-converting enzyme 2. *Circulation* **111**(20) 2605–2610.

[41] Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association* **94**(446) 496–509. https://doi.org/10.2307/2670170. MR1702320

[42] Fisher, R. A. (1925) *Statistical methods for research workers*. Oliver and Boyd. MR0346954

[43] Fitzmaurice, G. M. et al. (2008) *Applied Longitudinal Analysis*. Wiley.

[44] Fogelholm, M. and Kukkonen-Harjula, K. (2000). Does physical activity prevent weight gain–a systematic review. *Obesity reviews* **1**(2) 95–111.

[45] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*.

[46] Furno, P., Dionisi, M. S., Bucaneve, G., Menichetti, F. and Del Favero, A. (2000). Ceftriaxone versus β-lactams with antipseudomonal activity for empirical, combined antibiotic therapy in febrile neutropenia: a meta-analysis. *Supportive care in cancer* **8** 293–301.

[47] Garber, A. J., Duncan, T. G., Goodman, A. M., Mills, D. J., Rohlf, J. L. et al. (1997). Efficacy of metformin in

type II diabetes: results of a double-blind, placebo-controlled, dose-response trial. *The American journal of medicine* **103**(6) 491–497.

[48] Gelman, A. and Hill, J. (2013) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

[49] Ghosh, D. and Luo, S. (2025). A non-parametric U-statistic testing approach for multi-arm clinical trials with multivariate longitudinal data. *Journal of Multivariate Analysis* 105447. https://doi.org/10.1016/j.jmva.2025.105447. MR4901559

[50] Ghosh, D., Pal, S., Lutz, M., Luo, S. and Initiative, A. D. N. (2025). Ensemble survival analysis for preclinical cognitive decline prediction in Alzheimer's disease using longitudinal biomarkers. *Journal of Alzheimer's Disease* 13872877251365621.

[51] Ghosh, D., Xu, X., Luo, S. and Database, C. I. P. (2025). Power and sample size calculation for multivariate longitudinal trials using the longitudinal rank sum test. *Statistics in Medicine* **44**(20–22) 70261. https://doi.org/10.1002/sim.70261. MR4960437

[52] Ghosh, D., Boettcher, W. A., Johnston, R. and Lahiri, S. (2025). THANOS: A Predictive Model of Electoral Campaigns Using Twitter Data and Opinion Polls. *Data Science in Science* **4**(1) 2484180.

[53] Gibbons, J. D. and Chakraborti, S. (2010) *Nonparametric Statistical Inference*. CRC Press. MR2681063

[54] Gligorijevic, J., Gligorijevic, D., Pavlovski, M., Milkovits, E., Glass, L., Grier, K., Vankireddy, P. and Obradovic, Z. (2019). Optimizing clinical trials recruitment via deep learning. *Journal of the American Medical Informatics Association* **26**(11) 1195–1202.

[55] Gray, R. J. (1988). A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Annals of Statistics* **16**(3) 1141–1154. https://doi.org/10.1214/aos/1176350951. MR0959192

[56] Gueorguieva, R. and Krystal, J. H. (2004). Move over ANOVA: Progress in analyzing repeated-measures data and its reflection in papers published in the Archives of General Psychiatry. *Archives of General Psychiatry* **61**(3) 310–317.

[57] Guo, S., Jiang, X., Mao, B. and Li, Q. -X. (2019). The design, analysis and application of mouse clinical trials in oncology drug development. *BMC cancer* **19** 1–14.

[58] Halder, J. B., Benton, J., Julé, A. M., Guérin, P. J., Olliaro, P. L., Basáñez, M. -G. and Walker, M. (2017). Systematic review of studies generating individual participant data on the efficacy of drugs for treating soil-transmitted helminthiases and the case for data-sharing. *PLoS Neglected Tropical Diseases* **11**(10) 0006053.

[59] Harutyunyan, H., Khachatrian, H., Kale, D. and Ver Steeg, G. (2019). Multitask Learning and Benchmarking with Clinical Time-Series Data. *Scientific Reports*.

[60] Heinonen, M., Arora, S., Remes, S., Saarinen, I. and Lähdesmäki, H. (2021). Bayesian Multivariate Gaussian Processes for Longitudinal Clinical Data. *Journal of Biomedical Informatics* **114** 103654.

[61] Henderson, C. R. (1954). Estimation of variance and covariance components. *Biometrics* **9**(2) 226–252. https://doi.org/10.2307/3001853. MR0055650

[62] Hess, K. R. (1994). Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Statistics in medicine* **13**(10) 1045–1062. https://doi.org/10.1007/978-0-387-68639-4. MR2400249

[63] Ho, M. -W., Tu, W., Ghosh, P. and Tiwari, R. C. (2013). A nested Dirichlet process analysis of cluster randomized trial data with application in geriatric care assessment. *Journal of the American Statistical Association* **108**(501) 48–68. https://doi.org/10.1080/01621459.2012.734164. MR3174602

[64] Höfler, J., Rohracher, A., Kalss, G., Zimmermann, G., Dobesberger, J., Pilz, G., Leitinger, M., Kuchukhidze, G., Butz, K., Taylor, A. et al. (2016). (S)-Ketamine in refractory and super-refractory status epilepticus: a retrospective study. *CNS drugs* **30** 869–876.

[65] Hollander, M., Wolfe, D. A. and Chicken, E. (2013) *Nonparametric Statistical Methods*. Wiley. MR3221959

[66] Hong, J. and Chun, H. (2023). A prediction model for healthcare time-series data with a mixture of deep mixed effect models using Gaussian processes. *Biomedical Signal Processing and Control* **84** 104753.

[67] Hoo, J. -X., Yang, Y. -F., Tan, J. -Y., Yang, J., Yang, A. and Lim, L. -L. (2023). Impact of multicomponent integrated care on mortality and hospitalization after acute coronary syndrome: a systematic review and meta-analysis. *European Heart Journal-Quality of Care and Clinical Outcomes* **9**(3) 258–267.

[68] Huang, J. -Z., Chen, C. -N., Lee, C. -P., Kao, C. -H., Hsu, H. -C. and Chou, A. -K. (2022). Evaluation of the effects of skin-to-skin contact on newborn sucking, and breastfeeding abilities: a quasi-experimental study design. *Nutrients* **14**(9) 1846.

[69] Ibrahim, J. G. et al. (2001). Bayesian approaches to joint modeling of longitudinal and survival data. *Statistics in Medicine* **20**(13) 1993–2015.

[70] Ibrahim, J. G. and Molenberghs, G. (2009). Missing data methods in longitudinal studies: a review. *Test* **18**(1) 1–43. https://doi.org/10.1007/s11749-009-0138-x. MR2495958

[71] Izmailova, E. S., Wagner, J. A., Ammour, N., Amondikar, N., Bell-Vlasov, A., Berman, S., Bloomfield, D., Brady, L. S., Cai, X., Calle, R. A. et al. (2021). Remote digital monitoring for medical product development. *Clinical and Translational Science* **14**(1) 94–101.

[72] Kang, Q., Vahl, C. I., Fan, H., Geurden, T., Ameiss, K. A. and Taylor, L. P. (2019). Statistical analyses of chicken intestinal lesion scores in battery cage studies of anti-coccidial drugs. *Veterinary parasitology* **272** 83–94.

[73] Kay, R. (1977). The AFT model in survival analysis: Theory and applications. *Biometrics*.

[74] Ketema, T., Bacha, K., Getahun, K. and Bassat, Q. (2021). In vivo efficacy of anti-malarial drugs against clinical Plasmodium vivax malaria in Ethiopia: a systematic review and meta-analysis. *Malaria Journal* **20** 1–19.

[75] Khan, J., Ooka, J., Miller, S., Madden, L. and Hoitink, H. (2004). Systemic resistance induced by Trichoderma hamatum 382 in cucumber against Phytophthora crown rot and leaf blight. *Plant Disease* **88**(3) 280–286.

[76] Kim, A. Y., Jang, E. H., Kim, S., Choi, K. W., Jeon, H. J., Yu, H. Y. and Byun, S. (2018). Automatic detection of major depressive disorder using electrodermal activity. *Scientific reports* **8**(1) 17030.

[77] Kok, C., Jahmunah, V., Oh, S. L., Zhou, X., Gururajan, R., Tao, X., Cheong, K. H., Gururajan, R., Molinari, F. and Acharya, U. R. (2020). Automated prediction of sepsis using temporal convolutional network. *Computers in Biology and Medicine* **127** 103957.

[78] Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C. and Faisal, A. A. (2018). The Artificial Intelligence Clinician Learns Optimal Treatment Strategies for Sepsis in Intensive Care. *Nature Medicine* **24** 1716–1720.

[79] Koopmeiners, J. S. and Modiano, J. (2014). A bayesian adaptive phase i–ii clinical trial for evaluating efficacy and toxicity with delayed outcomes. *Clinical Trials* **11**(1) 38–48.

[80] Kumar, D. (2018) *Stress-Strength Estimation and its applications in Clinical Trials*. State University of New York at Albany. MR3908068

[81] Laird, N. and Ware, J. (1982). Random-Effects Models for Longitudinal Data. *Biometrics* **38** 963–974.

[82] Lehmann, E. L. and D'Abrera, H. J. M. (2006) *Nonparametrics: Statistical Methods Based on Ranks*. Springer. MR2279708

[83] Li, J., Chitwood, J., Menda, N., Mueller, L. and Hutton, S. F. (2018). Linkage between the I-3 gene for resistance to Fusarium wilt race 3 and increased sensitivity to bacterial spot

in tomato. *Theoretical and applied genetics* **131** 145–155.

[84] Li, L., Shen, C., Li, X. and Robins, J. M. (2013). On weighting approaches for missing data. *Statistical methods in medical research* **22**(1) 14–30. https://doi.org/10.1177/0962280211403597. MR3190643

[85] Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**(1) 13–22. https://doi.org/10.1093/biomet/73.1.13. MR0836430

[86] Little, R. J. and Rubin, D. B. (2019) *Statistical analysis with missing data*. John Wiley & Sons. https://doi.org/10.1002/9781119013563. MR1925014

[87] Luo, N., Di, W., Zhang, A., Wang, Y., Ding, M., Qi, W., Zhu, Y., Massing, M. W. and Fang, Y. (2012). A randomized, one-year clinical trial comparing the efficacy of topiramate, flunarizine, and a combination of flunarizine and topiramate in migraine prophylaxis. *Pain Medicine* **13**(1) 80–86.

[88] Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* **18**(1) 50–60. https://doi.org/10.1214/aoms/1177730491. MR0022058

[89] Mao, J. J., Bryl, K., Gillespie, E. F., Green, A., Hung, T. K., Baser, R., Panageas, K., Postow, M. A. and Daly, B. (2025). Randomized clinical trial of a digital integrative medicine intervention among patients undergoing active cancer treatment. *npj Digital Medicine* **8**(1) 29.

[90] Mashaly, O. A., El Mahallawy, A. S. and Amer, T. A. (2023). Intralesional Injection of Ethanolamine Oleate With or Without Local Anaesthetic Agent to Assess Postoperative Pain in Oral Venous Malformations (a Randomized Controlled Clinical Trial). *Alexandria Dental Journal* **48**(3) 102–108.

[91] Maxwell, S. E. and Delaney, H. D. (2004) *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Routledge.

[92] Mayan, I., Roth, H., Ghosh, D., Whitson, H. E. and Johnson, K. G. (2025). Genetic and biomarker disclosure process in a memory and aging study. *Journal of Alzheimer's Disease* **104**(2) 312–318.

[93] Mohamed, Y. R. E., El-Attar, A. M. I., Anwar, D. M. F. and Shehab, A. S. A. (2024). The efficacy of ultrasound and fluoroscopy-guided caudal epidural prolotherapy versus steroids for chronic pain management in failed back surgery syndrome. *Alexandria Journal of Medicine* **60**(1) 238–243.

[94] Molenberghs, G. and Kenward, M. (2007) *Missing data in clinical studies*. John Wiley & Sons.

[95] Nath, S., Korot, E., Fu, D. J., Zhang, G., Mishra, K., Lee, A. Y. and Keane, P. A. (2022). Reinforcement learning in ophthalmology: potential applications and challenges to implementation. *The Lancet Digital Health* **4**(9) 692–697.

[96] Nemati, S., Ghassemi, M. M. and Clifford, G. D. (2016). Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. In *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* 2978–2981. IEEE.

[97] Neuhaus, J. M. et al. (1991). Estimation of covariate effects in generalized linear models for longitudinal data. *Biometrics* **47**(4) 985–996.

[98] Noguchi, K., Gel, Y. R., Brunner, E. and Konietschke, F. (2012). nparLD: An R software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical Software* **50**(1) 1–23.

[99] Olson, C. L. (1976). On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*.

[100] Oyamada, S., Chiu, S. -W. and Yamaguchi, T. (2022). Comparison of statistical models for estimating intervention effects based on time-to-recurrent-event in stepped wedge cluster randomized trial using open cohort design. *BMC Medical Research Methodology* **22**(1) 123.

[101] Pan, C., Tian, Y., Zhou, T. and Li, J. (2024). Personalized Prediction of Parkinson's Disease Progression Based on Deep Gaussian Processes. In *MEDINFO 2023—The Future Is Accessible* 765–769 IOS Press.

[102] Park, Y. and Chang, W. (2024). A Personalized Dose-Finding Algorithm Based on Adaptive Gaussian Process Regression. *Pharmaceutical Statistics* **23**(6) 1181–1205.

[103] Pi-Sunyer, X., Astrup, A., Fujioka, K., Greenway, F., Halpern, A., Krempf, M., Lau, D. C., Le Roux, C. W., Violante Ortiz, R., Jensen, C. B. et al. (2015). A randomized, controlled trial of 3.0 mg of liraglutide in weight management. *New England Journal of Medicine* **373**(1) 11–22.

[104] Pinheiro, J. C. and Bates, D. M. (2000) *Mixed-Effects Models in S and S-PLUS*. Springer.

[105] Pocock, S. J. et al. (1987). Statistical considerations in the design of clinical trials: Beta-blockers in cardiovascular medicine. *Statistics in Medicine*.

[106] Pocock, S. J. et al. (2002). Subgroup analysis, covariate adjustment, and baseline comparisons in clinical trial reporting. *Statistics in Medicine*.

[107] President, T. (2019). Analyzing the Influence of Key Factors for Patient Willingness to Participate in Clinical Trials. *Semantic Scholar*.

[108] Proust-Lima, C. et al. (2009). Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research* **18**(2) 147–166.

[109] Psaty, B. M., Smith, N. L., Siscovick, D. S., Koepsell, T. D., Weiss, N. S., Heckbert, S. R., Lemaitre, R. N., Wagner, E. H. and Furberg, C. D. (1997). Health outcomes associated with antihypertensive therapies used as first-line agents: a systematic review and meta-analysis. *Jama* **277**(9) 739–745.

[110] Rasch, D. et al. (2011). The robustness of parametric statistical methods. *Psychological Science* **22**(9) 1211–1213.

[111] Rasmussen, C. E. and Williams, C. K. I. (2006) *Gaussian Processes for Machine Learning*. MIT Press. MR2514435

[112] Reitsma, A., Chu, R., Thorpe, J., McDonald, S., Thabane, L. and Hutton, E. (2014). Accounting for center in the Early External Cephalic Version trials: an empirical comparison of statistical methods to adjust for center in a multicenter trial with binary outcomes. *Trials* **15** 1–11.

[113] Ricciardi, F., Liverani, S. and Baio, G. (2023). Dirichlet process mixture models for regression discontinuity designs. *Statistical methods in medical research* **32**(1) 55–70. https://doi.org/10.1177/09622802221129044. MR4528435

[114] Rizopoulos, D. (2012) *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Chapman and Hall/CRC.

[115] Rogers, S., Farlow, M., Doody, R., Mohs, R., Friedhoff, L. and Group*, D. S. (1998). A 24-week, double-blind, placebo-controlled trial of donepezil in patients with Alzheimer's disease. *Neurology* **50**(1) 136–145.

[116] Rorden, C., Karnath, H. -O. and Bonilha, L. (2007). Improving lesion-symptom mapping. *Journal of cognitive neuroscience* **19**(7) 1081–1088.

[117] Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test. *Behavioral Ecology* **17**(4) 688–690.

[118] Schafer, J. L. (1997) *Analysis of incomplete multivariate data*. CRC press. https://doi.org/10.1201/9781439821862. MR1692799

[119] Scharf, A. -C., Gronewold, J., Eilers, A., Todica, O., Moenninghoff, C., Doeppner, T. R., de Haan, B., Bassetti, C. L. and Hermann, D. M. (2023). Depression and anxiety in acute ischemic stroke involving the anterior but not paramedian or inferolateral thalamus. *Frontiers in psychology* **14** 1218526.

[120] Schuler, M. S., Lechner, W. V., Carter, R. E. and Malcolm, R. (2009). Temporal and gender trends in concordance of urine drug screens and self-reported use in cocaine treatment studies. *Journal of addiction medicine* **3**(4) 211–217.

[121] Şenışık, S., Denerel, N., Köyağasıoğlu, O. and Tunç, S. (2021). The effect of isolation on athletes' mental health dur-

ing the COVID-19 pandemic. *The Physician and sportsmedicine* **49**(2) 187–193.

[122] Senn, S. (1993). Repeated measures ANOVA in clinical trials: Applications in weight loss and metabolic outcomes. *Biometrika*.

[123] Senn, S. (2006). Change from baseline and analysis of covariance revisited. *Statistics in Medicine*. https://doi.org/10.1002/sim.2682. MR2307596

[124] Shapiro, R. E., Hochstetler, H. M., Dennehy, E. B., Khanna, R., Doty, E. G., Berg, P. H. and Starling, A. J. (2019). Lasmiditan for acute treatment of migraine in patients with cardiovascular risk factors: post-hoc analysis of pooled results from 2 randomized, double-blind, placebo-controlled, phase 3 trials. *The journal of headache and pain* **20** 1–10.

[125] Sheller, M. J., Edwards, B., Reina, G. A., Martin, J. and Bakas, S. (2020). Federated Learning in Medicine: Facilitating Multi-Institutional Collaborations without Sharing Patient Data. *Scientific Reports*.

[126] Siddique, J. et al. (2008). Missing data in randomized controlled trials for weight loss. *Obesity*.

[127] Simoni, J. M., Wiebe, J. S., Sauceda, J. A., Huh, D., Sanchez, G., Longoria, V., Andres Bedoya, C. and Safren, S. A. (2013). A preliminary RCT of CBT-AD for adherence and depression among HIV-positive Latinos on the US-Mexico border: the Nuevo Dia study. *AIDS and Behavior* **17** 2816–2829.

[128] Sirima, S. B., Ouédraogo, A., Tiono, A. B., Kaboré, J. M., Bougouma, E. C., Ouattara, M. S., Kargougou, D., Diarra, A., Henry, N., Ouédraogo, I. N. et al. (2022). A randomized controlled trial showing safety and efficacy of a whole sporozoite vaccine against endemic malaria. *Science translational medicine* **14**(674) 3776.

[129] Song, S., Matsushima, N., Lee, J. and Mendell, J. (2015). Linear mixed-effects model of QTc prolongation for olmesartan medoxomil. *Journal of Clinical Pharmacology* **56**(1) 96.

[130] Sotoudeh-Paima, S., Segars, W. P., Ghosh, D., Luo, S., Samei, E. and Abadi, E. (2024). A systematic assessment and optimization of photon-counting CT for lung density quantifications. *Medical Physics* **51**(4) 2893–2904.

[131] Sousa, M. R. d. and Ribeiro, A. L. P. (2009). Systematic review and meta-analysis of diagnostic and prognostic studies: a tutorial. *Arquivos brasileiros de cardiologia* **92** 241–251.

[132] Student (1908). The probable error of a mean. *Biometrika* 1–25.

[133] Subbaswamy, A. and Saria, S. (2020). From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics* **21**(2) 345–352. https://doi.org/10.1093/biostatistics/kxz041. MR4132548

[134] Takada, M., Sozu, T. and Sato, T. (2015). Practical approaches for design and analysis of clinical trials of infertility treatments: crossover designs and the Mantel–Haenszel method are recommended. *Pharmaceutical statistics* **14**(3) 198–204.

[135] Takahashi, A. and Suzuki, T. (2021). Bayesian optimization design for dose-finding based on toxicity and efficacy outcomes in phase I/II clinical trials. *Pharmaceutical Statistics* **20**(3) 422–439.

[136] Teh, J. L., Purwin, T. J., Han, A., Chua, V., Patel, P., Baqai, U., Liao, C., Bechtel, N., Sato, T., Davies, M. A. et al. (2020). Metabolic adaptations to MEK and CDK4/6 cotargeting in uveal melanoma. *Molecular cancer therapeutics* **19**(8) 1719–1726.

[137] Therneau, T. M. and Grambsch, P. M. (2000) *Modeling Survival Data: Extending the Cox Model*. Springer. https://doi.org/10.1007/978-1-4757-3294-8. MR1774977

[138] Tian, W., Ding, W., Kim, S., Zheng, L., Zhang, L., Li, X., Gu, J., Zhang, L., Pan, M. and Chen, S. (2013). Efficacy and safety profile of combining vandetanib with chemotherapy in patients with advanced non-small cell lung cancer: a meta-analysis. *PLoS One* **8**(7) 67929.

[139] Trella, A. L., Zhang, K. W., Jajal, H., Nahum-Shani, I., Shetty, V., Doshi-Velez, F. and Murphy, S. A. (2024). A Deployed Online Reinforcement Learning Algorithm In An Oral Health Clinical Trial. *arXiv preprint arXiv:2409.02069*.

[140] Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica* **14**(3) 809–834. MR2087974

[141] Tsuboyama, K., Koyama-Honda, I., Sakamaki, Y., Koike, M., Morishita, H. and Mizushima, N. (2016). The ATG conjugation systems are important for degradation of the inner autophagosomal membrane. *Science* **354**(6315) 1036–1041.

[142] Tuomainen, K., Al-Samadi, A., Potdar, S., Turunen, L., Turunen, M., Karhemo, P. -R., Bergman, P., Risteli, M., Åström, P., Tiikkaja, R. et al. (2019). Human tumor–derived matrix improves the predictability of head and neck cancer drug testing. *Cancers* **12**(1) 92.

[143] Tushar, F. I., Vancoillie, L., McCabe, C., Kavuri, A., Dahal, L., Harrawood, B., Fryling, M., Zarei, M., Sotoudeh-Paima, S., Ho, F. C. et al. (2025). Virtual lung screening trial (VLST): An in silico study inspired by the national lung screening trial for lung cancer detection. *Medical Image Analysis* **103** 103576.

[144] Vangeneugden, T., Laenen, A., Geys, H., Renard, D. and Molenberghs, G. (2004). Applying linear mixed models to estimate reliability in clinical trial data with repeated measurements. *Controlled clinical trials* **25**(1) 13–30.

[145] Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*. Springer. https://doi.org/10.1007/978-1-4419-0300-6. MR1880596

[146] Vickers, A. J. (2001). The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient. *Trials*.

[147] Vinkers, D. J., Gussekloo, J., Stek, M. L., Westendorp, R. G. and Van Der Mast, R. C. (2004). The 15-item Geriatric Depression Scale (GDS-15) detects changes in depressive symptoms after a major negative life event. The Leiden 85-plus Study. *International journal of geriatric psychiatry* **19**(1) 80–84.

[148] Walker, M., Churcher, T. S. and Basáñez, M. -G. (2014). Models for measuring anthelmintic drug efficacy for parasitologists. *Trends in parasitology* **30**(11) 528–537.

[149] Wei, L. J. (1992). The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*.

[150] Welch, B. L. (1947). The generalization of 'STUDENT'S'problem when several different population varlances are involved. *Biometrika* **34**(1-2) 28–35. https://doi.org/10.2307/2332510. MR0019277

[151] Whitehead, J., Thygesen, H. and Whitehead, A. (2011). Bayesian procedures for phase I/II clinical trials investigating the safety and efficacy of drug combinations. *Statistics in Medicine* **30**(16) 1952–1970. https://doi.org/10.1002/sim.4267. MR2829058

[152] Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin* **1**(6) 80–83. Accessed 2024-12-01. https://doi.org/10.2307/3001946. MR0025133

[153] Wiles, N., Fischer, K., Cowen, P., Nutt, D., Peters, T., Lewis, G. and White, I. (2014). Allowing for non-adherence to treatment in a randomized controlled trial of two antidepressants (citalopram versus reboxetine): an example from the GENPOD trial. *Psychological medicine* **44**(13) 2855–2866.

[154] Wood, S. N. (2017) *Generalized additive models: an introduction with R*. Chapman and hall/CRC. MR2206355

[155] Wright, S. P. (1992). Adjusting for baseline in longitudinal clinical trials. *Journal of Clinical Epidemiology*.

[156] Xu, C., Hadjipantelis, P. Z. and Wang, J. -L. (2020). Semiparametric joint modeling of survival and longitudinal data: the r package JSM. *Journal of Statistical Software* **93** 1–29.

[157] Xu, R., Huang, S., Song, Z., Gao, Y. and Wu, J. (2024). A deep mixed-effects modeling approach for real-time monitor-

ing of metal additive manufacturing process. *IISE Transactions* **56**(9) 945–959.

[158] Xu, X., Ghosh, D., Luo, S. and Database, C. I. P. (2025). A novel longitudinal rank-sum test for multiple primary endpoints in clinical trials: Applications to neurodegenerative disorders. *Statistics in Biopharmaceutical Research* 1–11.

[159] Yakar, N., Emingil, G., Türedi, A., Şahin, Ç., Köse, T., Bostanci, N. and Silbereisen, A. (2023). Value of gingival crevicular fluid TREM-1, PGLYRP1, and IL-1β levels during menopause. *Journal of Periodontal Research* **58**(5) 1052–1060.

[160] Yoon, J., Jordon, J. and Schaar, M. (2018). Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning* 5689–5698. PMLR.

[161] Zeger, S. L. and Liang, K. Y. (1988). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **44**(4) 1049–1060. https://doi.org/10.2307/2532076. MR0999450

[162] Zhao, Y., Kosorok, M. R. and Zeng, D. (2009). Reinforcement learning design for cancer clinical trials. *Statistics in medicine* **28**(26) 3294–3315. https://doi.org/10.1002/sim.3720. MR2750277

[163] Zhao, Y., Zeng, D., Socinski, M. A. and Kosorok, M. R. (2011). Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics* **67**(4) 1422–1433. https://doi.org/10.1111/j.1541-0420.2011.01572.x. MR2872393

[164] Zhu, Y., Bi, D., Saunders, M. and Ji, Y. (2023). Prediction of chronic kidney disease progression using recurrent neural network and electronic health records. *Scientific Reports* **13**(1) 22091.

[165] Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology* **57**(1) 173–181. https://doi.org/10.1348/000711004849222. MR2087822

Dhrubajyoti Ghosh. Department of Biostatistics and Bioinformatics, Duke University, USA. E-mail address: dhrubajyoti.ghosh@duke.edu

Samhita Pal. Department of Statistics, North Carolina State University, USA. E-mail address: spal4@ncsu.edu