University of Helsinki
Dissertationes Universitatis Helsingiensis
128/2026

Advanced statistical estimation methods for event data and associated covariate processes in real-world settings

Ashwini Joshi

Academic dissertation

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public examination in Auditorium A111, Exactum, Kumpula, on the 1st of April 2026, at 13 o'clock.

Department of Mathematics and Statistics
Faculty of Science
University of Helsinki
2026

Supervisors Prof. Sangita Kulathinal
Department of Mathematics and Statistics
University of Helsinki, Finland

Prof. Dario Gasbarra
Department of Mathematics and Statistics,
University of Vaasa, Finland

Custos Prof. Sangita Kulathinal

Pre-examiners Prof. Jaakko Nevalainen
Faculty of Social Sciences (Biostatistics)
University of Tampere, Finland

Prof. (Retired) Uttara Naik-Nimbalkar
Department of Statistics
University of Pune, India

Opponent Dr. Amarjot Kaur
Department of Biostatistics and Research Decision Sciences
(Merck Research Labs)
Merck & Co., Inc., USA

Publisher: University of Helsinki
Series: Dissertationes Universitatis Helsingiensis 128/2026
ISBN 978-952-84-2127-6 (Paper back)
ISBN 978-952-84-2126-9 (PDF)
ISSN 2954-2898 (Print)
ISSN 2954-2952 (PDF)
PunaMusta, Joensuu 2026

Abstract

Event data, which are often encountered in medical studies, can be viewed and analysed in different ways based on the research questions of interest, the purpose of analyses, and the study designs employed for data collection.

The research articles in this thesis discuss event data in three different settings: event data at a fixed time or in a fixed time interval, multiple types of event at multiple follow-up times in a longitudinal setting, and event data in a time-to-event setting. The objectives of this thesis are understanding the limitations of applicability of classical statistical methods to real-world event data, identifying possible biases in classical methods, and suggesting improvements in existing methods in various event data settings. When events are counted in some fixed time period, then it is treated as count data. The Poisson distribution is a natural choice for studying count data. However, based on the research question and the setting, binomial or negative-binomial distribution and corresponding regression methods can also be used. These regression models fall under the umbrella of generalised linear models (GLMs). GLMs, like many other classical statistical methods, may not give a valid inference when the sample size is small, a problem generally known as small-sample bias. This thesis discusses frequentist parametric (Firth penalisation method), semiparametric (Exact permutation method), and Bayesian data augmentation methods of bias-reduction for parameter estimation and their effect on event prediction. All alternative methods are evaluated for their performance in estimation and prediction in a comprehensive simulation study under a variety of realistic scenarios. The main conclusion of this research was that two modifications of Firth's approach and the Bayesian estimator are universally applicable solutions providing bias reduced parameter estimation and bias reduced event prediction.

Small-sample bias intensifies if there are multiple types of event. When multiple disease markers are present, multiple types of events can be defined related to them. In this thesis, an existing Bayesian method for a single marker is extended to multiple disease markers to obtain simultaneous inference. Joint posterior distribution of multiple markers is studied with special attention to posterior credible regions. The application of equal tailed credible regions and highest posterior density credible regions to gatekeeping procedures is developed to apply a hierarchy of markers. Simulation studies with many combinations of prior information and data generating distributions were carried out to examine frequentist properties of these simultaneous credible regions. When the treatment response is captured by multiple markers, studying each marker separately is not sufficient to understand the treatment response since the dependency among multiple markers

is not accounted for. Methods that give combined inference such as simultaneous credible regions are useful in such cases as illustrated in the simulation study and a case-study of multiple markers data of the wet type of age-related macular degeneration of the eye (wet-AMD).

In the third article, nonparametric methods for jointly modelling longitudinal follow-up data involving measurements of multiple disease markers or covariates along with the time to an event of interest are discussed. Examining the history of time-dependent covariates in longitudinal settings and their effect on the event of interest can help in deciding future treatment. The integrated covariate process is a good measure of the cumulative response over time. In the third article of this thesis, there have been major methodological advances in the estimation of bivariate survival functions in the presence of a special type of censoring mechanism of the bivariate event times. In order to handle censoring and to use all the available information, inverse-probability of censoring weighted estimators of survival probability are developed. The estimators and their efficiency differ according to the use of the information in estimating the censoring distribution. Understanding the contribution of each individual in the joint survival estimation is the strength of this article.

Data from clinical visits of patients suffering from wet-AMD are used to illustrate the usefulness of the methods discussed in this thesis. All articles in this thesis discuss improvements in standard methods and also propose new methods for real-world medical data in order to use the available information in an efficient way and to facilitate medical decision making.

Acknowledgments

My work on event data started with the purpose of development of software for Firth penalisation to GLMs during my employment at Cytel statistical software and Services, Pune, India. My mentor at Cytel, Dr. Pralay Senchaudhuri motivated me to pursue further research in this area. I co-authored the first article of this thesis with Pralay and other collaborators of Cytel. I am grateful to Dr. Pralay Senchaudhuri and Prof. Georg Heinze for encouraging me to develop varied types of methods for the small sample bias problem. Around that time, I received a mobility grant from the Academy of Finland to work with Prof. Sangita Kulathinal at the University of Helsinki on a research problem related to age-related macular degeneration of the eye. In next few years, Sangita introduced me to varied statistical study areas from fully theoretical to fully application oriented. Event history analysis was the gem of it. During my PhD journey, Sangita has been my mentor and supervisor. Without her this journey would have been impossible. My other supervisor Prof. Dario Gasbarra introduced me to various ways to ensure mathematical accuracy of the methods we developed. I am grateful to both my supervisors for insisting on critical thinking and asking crucial questions which encouraged me to think deeper.

Getting the correct and clean data is half the battle won for a statistician. I owe my gratitude to my coauthors and collaborators from the department of ophthalmology, Helsinki University Hospital, Adjunct Prof. Dr. Ilkka Immonen and Dr. Terhi Ollila. They provided the data and their medical knowledge about age-related macular degeneration. I am also thankful to all my other collaborators and coauthors - Dr. Isha Dewan, Prof. Christopher Corcoran, Dr. Angelika Geroldinger, Lena Jiricka, and Juuso Silvennoinen. I am thankful to my thesis committee members, Dr. Leena Kalliovirta and Prof. Samuli Siltanen for regular reviews and encouragement.

I am very grateful to Freya Barua for providing the language editing help on a tight schedule. Also thanks to Dr. Leena Kalliovirta for the Finnish translation of the short abstract. I am very grateful to the two pre-examiners, Prof. Jaakko Nevalainen and Prof. (Retd.) Uttara Naik-Nimbalkar, who took the time and effort to review my thesis and provided valuable suggestions for improvement. I am most grateful to Dr. Amarjot Kaur who accepted to be my opponent for my thesis defence.

During my PhD, I worked full-time at Akkurate Oy (initially and then Sandvik mining and Construction Oy). I am grateful to my supervisors and team members for supporting me by scheduling my work to help me attend important statistical lectures, seminars, and discussions. Special thanks to my manager Jari Koskinen

for facilitating my study leave which enabled me to complete the final parts of my thesis.
Special thanks to Sangita, Bijoy and Xabi Kulathinal for their support throughout this journey - from buying groceries during the first few days of quarantine to introducing us to the life and the customs in Finland to taking us on some nature trips was all incredible.
My journey to Finland started with my then 12-year old daughter during the Covid era, which was a difficult time for everyone. It would have been hard to continue without the support from family and friends. I am thankful to my parents, in-laws and extended family for all the support and encouragement they extended towards me. My husband, Shrikant, encouraged me to fulfil my dream while he travelled back and forth between India and Finland to fulfil our duties in my absence. My daughter, Irawati, showed great courage in accompanying me to an unknown land and showed great endurance throughout all these years.
I would like to thank the Academy of Finland for the research mobility grant (number 33137) and Alfred Kordelinin Säätiö for the thesis completion grant.

Ashwini Joshi

The thesis consists of this summary and the following original research articles referenced using Roman numerals **I** – **III**.

List of original publications

I **Joshi, A.**, Geroldinger, A., Jiricka, L., Senchaudhuri, P., Corcoran C., Heinze G. (2021). Solutions to problems of nonexistence of parameter estimates and sparse data bias in Poisson regression. Statistical Methods in Medical Research. 31(2):253-266. https://doi.org/10.1177/09622802211065405

II **Joshi, A.**, Ollila, T., Immonen, I., and Kulathinal, S. (2022). Bayesian Simultaneous Credible Intervals for Effect Measures from Multiple Markers. Statistics in Biopharmaceutical Research, 15(3), 638–653. https://doi.org/10.1080/19466315.2022.2110935

III **Joshi, A.**, Gasbarra, D., Kulathinal, S. (2025). Nonparametric estimation of the joint and conditional survival functions of the time to an event of interest and associated integrated covariate processes. (under review). Pre-print: Authorea. December 04, 2025. https://doi.org/10.22541/au.175822739.96310819/v2

Author's contributions

I First author AJ alongwith PS, GH started the project with conceptualisation, simulation study and an initial draft. AJ developed statistical methods and simulations in R and LogXact. GH developed statistical methods in SAS. AG contributed in writing successive drafts. LJ contributed in better graphical representation of results. CC and PS provided overall statistical guidance and language improvement. GH provided expertise in Firth correction methods for MLE in all aspects of developing new methods, writing and choosing appropriate datasets.

II AJ and SK started the project with conceptualisation, simulation study and an initial draft. AJ developed statistical methods and simulations in R and wrote all drafts. TO and II provided the data on wet-AMD patients and the background knowledge for wet-AMD disease. SK provided overall statistical guidance in all aspects of developing new methods, writing and choosing appropriate datasets.

III AJ and SK started developing the statistical method for joint survival function and writing the initial draft. DG introduced the conditional hazard function theory and provided guidance on theoretical accuracy of results. AJ performed all computations, graphical representation and wrote all drafts. SK provided overall statistical guidance in all aspects of developing new methods, writing and choosing appropriate datasets.

Contents

Abbreviation **x**

1 Introduction **1**
1.1 Datasets . . . 3
1.2 A short note on wet-AMD disease . . . 7
1.3 Objectives . . . 9

2 Rare events, Poisson regression, and penalisation **10**
2.1 Non-existence of ML estimates and alternative methods . . . 12
2.2 Bias reduction in prediction . . . 16

3 Comparisons of groups, effect measures and multiple markers **19**
3.1 Contingency tables and simultaneous credible intervals . . . 22
3.1.1 Gatekeeping . . . 24
3.2 Longitudinal measurements of multiple markers . . . 26
3.2.1 Extension to time-dependent contingency tables . . . 27
3.2.2 Clustering of trajectories . . . 28

4 Time-to-event and associated covariate processes **35**
4.1 Nonparametric estimation . . . 37
4.1.1 Statistical formulation . . . 37
4.1.2 Joint survival estimation . . . 40
4.2 Medical decision making . . . 48
4.2.1 Role of conditional hazards and estimation . . . 48
4.3 Comments on proposed estimators . . . 50

5 Article summaries **52**
5.1 Article **I** . . . 52
5.2 Article **II** . . . 53
5.3 Article **III** . . . 54

6 Discussion **56**

6.1 Variance computation and interval estimation 57
6.2 Computation of the area under the trajectory 59
6.3 Real-world clinical data and biases 60
6.4 Possible extensions 61
6.5 Concluding remarks 65

Abbreviations

- ML: Maximum likelihood estimation method
- GLM: Generalised linear model
- MCLE: Maximum conditional likelihood estimate
- MUE: Median unbiased estimate
- PL or FL: Penalised likelihood or Firth likelihood
- FLIC: FL with intercept correction
- FLAC: FL with added covariate
- ETI: Equal tailed interval
- HDI: Highest density interval
- AMD: Age-related macular degeneration
- wet-AMD: wet type of AMD in which excess fluid on retina blurs the central vision
- VA: Visual acuity of an eye
- CFT: Central foveal thickness of an eye
- IRF or cyst: Intra-retinal fluid
- SRF or NED: Sub-retinal fluid, abnormal fluid between the retina and the retinal pigment epithelium
- PED: Pigment epithelial detachment
- CNV lesion: A contiguous area of abnormal tissue in the macula that encompasses angiographically documented CNV with possible additional components of subretinal haemorrhage, blocked fluorescence not from haemorrhage, serous detachment of the retinal pigment epithelium, and fibrosis.

- FA: Fluorescein Angiography
- OCT: Optical Coherence Tomography
- Active CNV: includes both, leakage on FA and (SRF or IRF) fluid on OCT.
- KM: Kaplan-Meier estimator of a survival function
- IPCW: Inverse-probability-of-censoring-weighted estimation method.

Chapter 1

Introduction

The data in medical studies often contain various types of event data. In this thesis, innovative statistical methods are developed or existing methods are carefully modified for three types of event data, based on how the information about the concerned time period is viewed. In this thesis, the terms events, outcomes, responses, markers are used interchangeably. Event data can be considered as count data, where the events occurring within a fixed time period are counted, viz. the number of hospitalisations. Such counts are generally studied using Poisson distribution (Cameron and Trivedi, 1998; Jones et al., 2013). In some studies, event rate is the parameter of interest. Poisson distribution is useful in this case if the assumption of the equality of mean and variance is valid (Frome, 1983). When the actual time of the event is of interest, then it has to be modelled. In this situation, Poisson distribution is still useful and it corresponds to a constant event rate over time (Perperoglou, 2011). If the variance is much larger than the mean, which indicates an overdispersion in the data, then some alternative to Poisson distribution is needed (Berk and MacDonald, 2008; Hoef and Boveng, 2007). For overdispersed data, negative-binomial distribution is commonly used. When the experiment continues until the pre-specified events are observed, the data are generally skewed and overdispersed. In such cases, negative binomial distribution is the natural choice (Hilbe, 2011). When individual-specific covariates are available, the binary status of observing an event can be modelled using logistic regression, where the event probability depends on the covariates (Agresti, 2013). When multiple observations over a follow-up period are available, the methods for longitudinal data are useful (Fitzmaurice et al., 2011). Lastly, a very important type of event data comprises those involving the time to a specific event. When the interest is in the study of the time to an event then survival analysis comes to the rescue (Kalbfleisch and Prentice, 2002).

In this thesis, statistical methods for estimation are developed for the event

data in most of the above mentioned scenarios. Most of the motivating examples are from medical observational studies, however similar datasets from randomised clinical trials are also used for demonstrating the use of proposed methods. There are many standard classical methods available for large sample studies or randomised studies involving event data. In randomised clinical trials, generally, the difference between two treatments is studied and the sample size required to establish the difference between the two treatments is decided based on the analysis method to be used. Strict inclusion and exclusion criteria are in place to avoid any effect of confounding variables. Observational studies have different objectives than clinical trials. Clinical trials are designed to study the usefulness of a new treatment. So, the focus is on proving the difference between the new treatment and the existing treatment, or between the new treatment and a placebo. In clinical practice, however, the objective can be to study the effect of an already approved treatment for subgroups of patients with specific characteristics, or to choose the most suitable treatment from many available treatments for a patient, according to the patient's medical condition. Further, patients treated in clinical practice may have many comorbidities, medical conditions which are not all studied in a clinical trial. These are some differences in clinical trials and observational studies which make it difficult to use the same statistical analysis methods in both situations. Another reason of the non-applicability of the same methods in these two situations is that the assumptions of classical methods are generally not met in observational studies. For example, small sample studies or rare events studies do not have a sufficient sample size or sufficient events, which is a hurdle while applying many classical methods. Also, individuals in observational studies are not randomised to different treatments, and hence, any direct comparisons of treatment effects between different treatments may be difficult to interpret. Due to such differences in the objectives and the settings of the two types of studies, the statistical methods used for analyses cannot be applied directly and require modifications. Moreover, there are other general problems of biases, especially in observational studies. The potential biases are associated with the study design as well as with the specific type of analyses. The most relevant biases in the context of the case studies of the thesis are discussed in section 6.3.

For the datasets that may not obey the assumptions of large sample methods, various alternative methods are developed and discussed in this thesis. These alternative methods give less biased results, as compared to large sample methods, but they have their own issues, like computational complexity, and a shifting of bias to some other parameters that are not of interest. These methods use more involved statistics, which makes it difficult for a non-statistician to apply them.

Table 1.1: Publication highlights

Article	Research questions	Approaches	Dataset type
I	Bias in small sample estimation and prediction	Parametric: Penalised likelihood, Semiparametric: Exact and mid-p, Bayesian: Jeffreys prior	Number of events in the fixed time interval
II	Multiple outcome-simultaneous inference, hierarchical inference, different effect measures	Bayesian framework for multiple effect measures, Simultaneous credible intervals, Extending Gatekeeping	Multiple types of events at two time points or in two groups
III	Survival analysis and decision making based on integrated covariate processes	Nonparametric - Joint and conditional survival	Longitudinal and survival data with independent right censoring

Also, the assumptions underlying these methods are often overlooked, which may lead to their misuse or to misinterpretation of their results.

In this thesis, all three research articles discuss such alternative methods for event data, along with their advantages and their limitations in detail. Table 1.1 lists the methodologies and the datasets discussed in this thesis.

All three articles discuss either nonparametric or Bayesian methods that deviate from the standard methods to address the problem at hand and to deal with the event data in various formats. The methodologies developed in all three articles are helpful in making policy decisions.

1.1 Datasets

The details of the datasets used in this thesis are given below.

Covid data : These were rare events data of coronavirus infections used in Article **I**. During the outbreak of the coronavirus disease (COVID-19) in the winter of 2020 and the spring of 2021, the employees of supermarkets, nursing homes, and hospitals were considered as key professionals, since they were in contact with many clients daily, even under the lockdowns of public life that were imposed by many countries at that time. Many of their

clients, such as elderly people or people with chronic diseases, were considered to be at a high risk of a severe course of the disease if they got infected. At some point, the question arose whether more stringent controls for virus spread should be imposed on one of these groups, in particular, considering the unknown probability of asymptomatic infections. Therefore, on 28 March 2020, and on 30 March 2020, a representative series of 1161 tests for the presence of an infection with the novel coronavirus was performed among randomly selected asymptomatic employees of supermarkets, nursing homes, and hospitals in Austria. Out of 1161 persons tested, only six tested positive, with three each in the groups of the employees of nursing homes and the employees of hospitals, and none among the employees of supermarkets. The research question behind these tests was how the risk of infection differs between the employees of supermarkets, the employees of nursing homes, and the employees of hospitals. The question could be answered by using Poisson regression by comparing the risk ratios of nursing homes and hospitals to those of supermarkets. These data were published in an Austrian newspaper (Al-Youssef et al., 2020). We chose these data for demonstration of usefulness of our method because the data had zero event in one group, making the comparison of the two groups difficult and suggests the need for new methods for sparse data. The data are discussed in more detail in Article **I**. Also, such research questions have a greater impact on policy decisions.

Implant dentistry data : Another set of rare events data were used in Article **I**. The data were about the risk factors in the complications in implant dentistry (Feher et al., 2020). The risk factors were assessed in 1133 patients undergoing 2405 implantations. The haematological complications by the risk factors, including the patient's age (in decades), the patient's smoking habits (no smoking, light smoking, or heavy smoking), the presence of diabetes mellitus, and the number of implantations performed per patient, were available in the data. In Poisson regression, the logarithm of the number of implantations was considered as the offset, and the number of complications as events for each patient. We chose these data because no haematological complications were observed for light smokers, making this group of patients an ideal candidate for studying the sparse data. Drs Kuchler, Gruber, and Feher from the University Clinic of Dentistry, Medical University of Vienna provided the data.

Migraine clinical trial data : These data were used in Article **II** to demonstrate the use of our methodology of handling multiple markers simultane-

ously for clinical trial data. These data were a small subset of the data from a clinical trial (Ho et al., 2008). We considered two groups of 50 patients each, with one group on the treatment (Telcagepant), and the other one on placebo. These data were taken from the software StatXact (version 12) ("StatXact 12, Cytel Inc. Cambridge, MA. Software Manual", 2019). Here, we considered four binary markers: pain freedom at two hours post treatment (EP1), absence of phonophobia (sensitivity to sound) at two hours post treatment (EP2), absence of photophobia (sensitivity to light) at two hours post treatment (EP3), and sustained pain freedom up to a period of 24 hours (EP4). These data had multiple binary outcomes or endpoints which made it suitable for testing the methods developed in Article **II**. Further, as the data were analysed previously by other methods, we could compare our new methods against the earlier results.

wet-AMD real-world data : These data of age-related macular degeneration of the eye (AMD) was the motivation for articles **II** and **III**. A brief note on the wet-AMD disease is given in section 1.2. We considered the data of 110 patients suffering from the wet type of age-related macular degeneration (wet-AMD). These data were made available by our collaborators from Helsinki University Hospital (HUS). A general description of this cohort and the data from the medical perspective is available in (Ollila et al., 2022). Here a short description of the data is given which is necessary for statistical methodology development. The wet-AMD treatment involved anti-VEGF injections that control the bleeding inside the eye by stopping the growth of new blood vessels. Treatment decisions had been made by clinicians in routine practice mostly by using either the *PRN* (injection when disease activity is seen) protocol or the *treat-and-extend* (a treatment interval that minimises the disease activity) protocol. Typically, the visit interval was 4 to 5 weeks. The baseline measurements and the follow-up measurements of up to four years of treatment were available. The measurements included visual acuity (VA) measured on a continuous scale at every monthly visit, and other lesions measured less frequently, like Central retinal thickness (CFT) on a continuous scale, and the binary status (presence/absence) of abnormal fluid within the retina as IRF, SRF, and PED. The actual fluid volumes were not available. The value of CFT lying roughly between 200 and 300 is generally considered as a healthy value, although there is no consensus on this number among the clinicians. The markers IRF, SRF, and PED were binary, while the markers VA and CFT were continuous. Visual acuity was measured every month, since it is the main criterion of

vision quality and is easy to measure. Other markers were measured less frequently, since their measurements can be inconvenient for the patients, besides being expensive. The value of CFT outside the healthy range or the presence of abnormal fluid are the main indicators of the disease activity. The decision of administering an injection is guided by these markers. Although they are not measured frequently, their effect on the vision can be examined from the regular measurements of VA. It is clear that the accumulation of fluid in the retina causes vision problems. However, it is not clear whether the complete drying of the fluid or the presence of some amount of fluid helps in better vision. For wet-AMD IRF, SRF, PED, and CFT are considered the anatomical markers, while VA is the physical marker of the disease. All these markers are needed for understanding the disease activity. Considerable statistical challenges arose because these anatomical markers were not measured at all visits and the volume of the fluid was not available. In Article **II**, all these markers in dichotomised form are used for demonstrating the proposed method of joint credible regions of multiple markers. In Article **III**, VA trajectory is used for computing the area under the trajectory and the discontinuation of treatment is considered as the event of interest.

wet-AMD clinical trial data : These data are used in Article **III**. These data are from the CATT (Comparison of Age-Related Macular Degeneration Treatments Trials) clinical trial study. The CATT data consist of a 2-year follow-up of 1185 patients suffering from wet-AMD. More details of this study are available at https://www.med.upenn.edu/cpob/catt.html. Only a part of the data are used here for illustration purposes. We considered two subgroups of the data, based on the treatment regime: regular injection intervals with *intent-to-treat* (injection at pre-decided times) protocol or variable injection intervals with *PRN* (injection when disease activity is seen). In the group with regular injection intervals, each patient received an injection every four weeks irrespective of the status of the other lesions (leakage on fluorescein angiography and subretinal, intraretinal, or sub-RPE fluid on OCT), while the patients in the group with variable injection intervals received an injection only if the patient showed signs of higher disease activity through the other lesions. All patients visited the clinic for measuring the visual acuity every four weeks. Each patient remained in the same treatment group throughout. For the regular group, since the treatment did not depend on the lesion measurements, fluid status was not measured in all visits. For the variable group, the decision to administer

an injection depended on both, VA and lesion status. So, the fluid status was available at each visit. The VA trajectory of a randomly chosen patient from each group is shown in Figure 1.1. In Article **III**, this data was used for demonstrating the use of proposed method in comparing two groups of patients. VA trajectory is used for computing area under the trajectory and the discontinuation of treatment is considered as the event of interest.

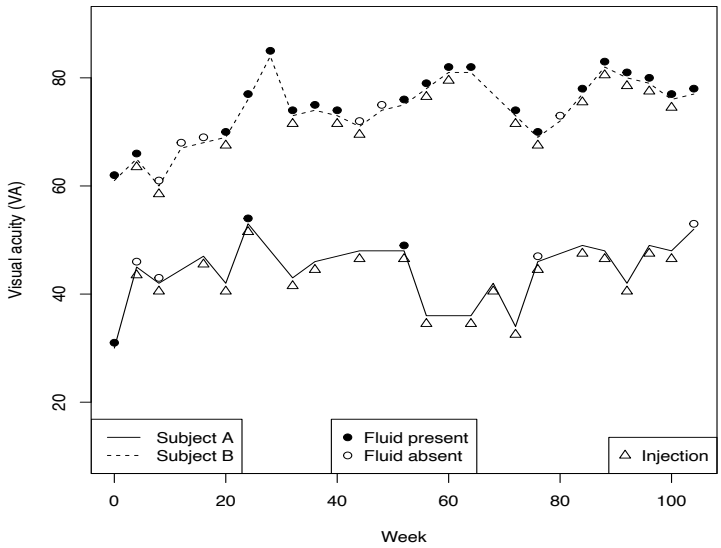


Figure 1.1: Visual acuity (VA) trajectories of two patients of the CATT study. The trajectories are obtained by connecting consecutive visual acuity measurements by lines. Visual acuity trajectories for two individuals, one each from the regular group (solid line) and the variable group (dashed line) of the CATT study data are plotted. If fluid measurements are available for a visit, the fluid status is shown by a circle above the trajectory, a filled circle indicates the presence of fluid and an empty circle indicates the absence of fluid. If an injection is administered at a visit then it is shown by a triangle below the trajectory. In both groups, VA was measured at all visits if the patient's physical condition allowed it. For individual A in the regular group, all the visits show an injection as per the regular injection regime, while for individual B, an injection was administered only when there was a higher disease activity measured through the other lesions including the fluid.

1.2 A short note on wet-AMD disease

Articles **II** and **III** were motivated by the age-related macular degeneration of the eye disease. Hence, a brief description of the disease and its treatment are pre-

sented here. Age-related macular degeneration (AMD) of the eye, especially its variant wet-AMD, is a chronic disease, which is one of the leading causes of visual handicap in Finland (Tuuminen et al., 2017) and also worldwide (Flaxman et al., 2017; Nair et al., 2022). In this disease, A patient with AMD starts losing eyesight from the central region of the retina. The most severe and long-standing handicap in AMD is mediated by the secondary growth of new vessels under the retina (wet-AMD). The anti-vascular endothelial growth factor (anti-VEGF) therapy, in which anti-VEGF agents are injected intra-ocularly and repeatedly for years, inhibits new vessel growth under the retina (Chakravarthy et al., 2010). There is a variability in the biological effectiveness and the costs of different anti-VEGF agents available. Moreover, new agents with different biological targets are being introduced (Hussain et al., 2019). Currently, the treatment protocol used by most centres is of the type 'treat-and-extend', in which the injection frequency is adjusted individually to find the longest injection interval that keeps the signs of retina activity absent (Gillies et al., 2015). A detailed analysis of AMD outcomes, especially in real-world settings, is important for optimizing treatment strategies, health economics, evaluating new drugs, pharmacogenetic outcomes, and benchmarking. Real-world AMD materials consist of multiple variations of activity markers, treatment intervals and medications. In addition to this, dropouts are common. This diversity makes the evaluation and the comparison of the treatments difficult. The response to the treatment varies between individuals and drugs used. Visual acuity gives a direct measure of the vision in the measured eye and is easy to measure. Hence, it can be taken as the primary endpoint and can be measured more frequently as compared to the other endpoints, like lesions and subretinal fluids. The impact of these other outcomes on visual acuity and, in turn, on the overall vision is not yet well understood. There are at least four markers of lesion activity and the inter-individual temporal variation in all of these is observed. Moreover, these outcomes are of different types, such as continuous, categorical, binary. Such multivariate mixed type of data make statistical modelling and inference complex. Also, there is a suboptimal treatment visit attendance and a significant patient dropout during the treatments. Thus, until now, most reports in this field are based on mean visual acuity values at fixed, often annual, time points and a mean number of injections. The impact of patients dropping out may not have been dealt with at all in many studies.

1.3 Objectives

The objectives of this thesis are to either modify or develop statistical methods addressing bias in maximum likelihood estimation and prediction due to small sample or rare events, enabling comparisons of groups with respect to multiple binary markers data, and facilitating medical decision making resulting from the complex interplay between the time to an event of interest and associated marker or covariate processes. The statistical methodological research in this thesis has its origin in one or more clinical studies. The specific objectives are:

1. To address issues of bias in maximum likelihood estimation and prediction in Poisson regression models for small sample or sparse event. [Article **I**]

2. To compare two groups of patients with respect to multiple binary markers using various effect measures. [Article **II**]

3. To estimate the joint survival probabilities of the time to an event of interest and the stopping times defined using the associated covariate processes in the presence of a special type of censoring mechanism. [Article **III**]

The rest of this thesis is structured as follows. In Chapters 2, 3, and 4, I describe the necessary statistical background and discuss the methodologies developed in the three articles. In order to achieve the first objective, in Article **I**, we reviewed some existing bias reduction methods available for maximum likelihood method of parameter estimation in Poisson regression models. We modified some of these methods to obtain bias reduced predictions of events. I discuss these methods in more detail in Chapter 2. In Chapter 3, I elaborate on statistical methods related to multiple markers. A Bayesian method of simultaneously credible intervals to obtain simultaneous inference from multiple binary markers was developed in Article **II**. I also discuss methods of clustering longitudinal trajectories of multiple markers. In Chapter 4, I discuss statistical methods developed in Article **III** to understand the role of covariate processes associated with an event of discontinuation of treatment in the real-world setting. In Chapter 5, a brief summary of the three research articles is given. In Chapter 6, I describe the main findings of this research and the future directions.

Chapter 2

Rare events, Poisson regression, and penalisation

The number of events of a certain type at a fixed time or during a fixed interval is the focus of this chapter. The number of events during a fixed interval is considered as count data and is often modelled using Poisson distribution. When there are one or more independent covariates, and the response variable is the number of events, to understand the association between the response and the covariates, Poisson regression is employed. The regression coefficients concisely describe this association, and are often the parameters of interest. For large datasets, the standard frequentist methods like maximum likelihood (ML) estimation are used for parameter estimation. When data samples are small, or events are rare, such methods may not give a valid inference. For small sample or rare events data, maximum likelihood estimation is known to be biased. This bias is studied in the literature and bias expressions are given by many authors (Cox and Hinkley, 1974; Mardia et al., 1999; McCullagh and Nelder, 1989). By using the ML estimate (MLE) in place of the true parameter value in these expressions or by using the jackknife method, this bias can be estimated. In these methods, MLE is computed as a first step and then the estimated bias is subtracted from it to get a bias reduced parameter estimate. These methods are useful when finite MLE exists but in case of non-finite MLE, they cannot be used. Firth (1993) suggested a method which did not require MLE computation as the first step. We discussed this method at great length in Article **I**. Firth's penalisation method reduces the small sample bias and shrinks parameter estimates toward zero. This method is especially useful when the ML estimate is non-finite, since it gives a finite estimate by penalisation. The ML method gives non-finite estimates for regression coefficients in small sample studies when there is separation or quasi-separation in the data. When a covariate or a group of

covariates determines seemingly perfectly the response of observing or not observing an event, then a complete separation is observed. For small sample or rare-events data, the possibility of separation or quasi-separation is high, as seen in Covid data of Covid infections in Article **I** where one of the group has zero events. The cases of separation and quasi-separation are illustrated for a binary response in Figure 2.1. This figure is taken from (van Smeden et al., 2016) under license (http://creativecommons.org/licenses/by/4.0/).

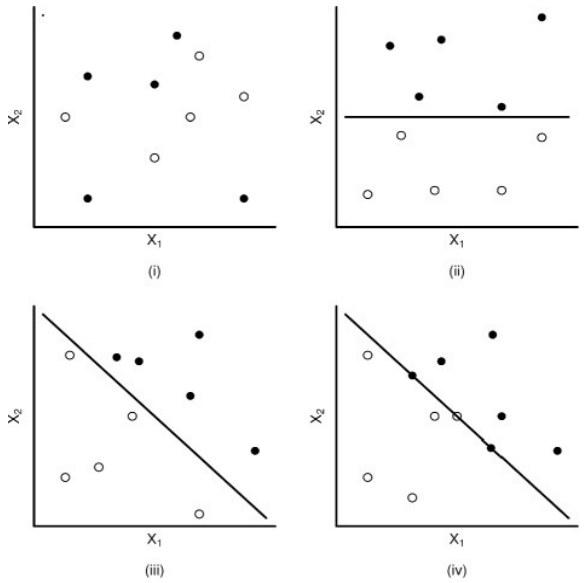


Figure 2.1: Graphical representation of separation (complete and quasi-complete). Different scenarios of overlap or cases of separation are shown with scatter plots of 2 covariates. The filled and the empty circles represent the levels of a binary response. The top-left plot shows an overlap, the top-right plot shows a complete separation due to the variable x_2 but not involving x_1. The lower-left plot shows a complete separation of the response with a linear combination of x_1 and x_2. The lower-right plot shows a quasi-separation of the response with a linear combination of x_1 and x_2. This figure is taken from (van Smeden et al., 2016) with slight modification and a new caption under license (http://creativecommons.org/licenses/by/4.0/).

Separation in the case of a binary response in logistic regression is easy to detect and has been studied in the literature (Albert and Anderson, 1984; Heinze and Schemper, 2002; Puhr et al., 2017). Other regression models in the family of generalised linear models (GLM) also suffer from the non-existence of parameter estimates due to separation. A necessary and sufficient condition for the non-existence of ML estimates for a wide class of generalised linear models is identified in Correia et al. (2019). The Poisson distribution falls within the exponential family of distributions and the Poisson log-linear regression model is a special case

of GLM. For Poisson regression, the condition of non-existence of ML estimates can be understood geometrically. The data are separated if and only if all data points corresponding to the non-zero events and all data points corresponding to zero events are on the opposite sides of some hyperplane of the data. If a hyperplane separates non-zero events and zero events but some of the non-zero events and some of the zero events lie on the hyperplane, it corresponds to a quasi-separation in the data. We discussed the condition for the non-existence of ML estimates and its proof for Poisson regression in Article **I**. For datasets with a complete separation or quasi-separation, like Covid data, the maximum likelihood estimate does not exist, and alternative methods are necessary. I discuss some of them in the next section.

2.1 Non-existence of ML estimates and alternative methods

We studied many alternative methods to overcome the problem of non-existence of ML estimates in Article **I**.

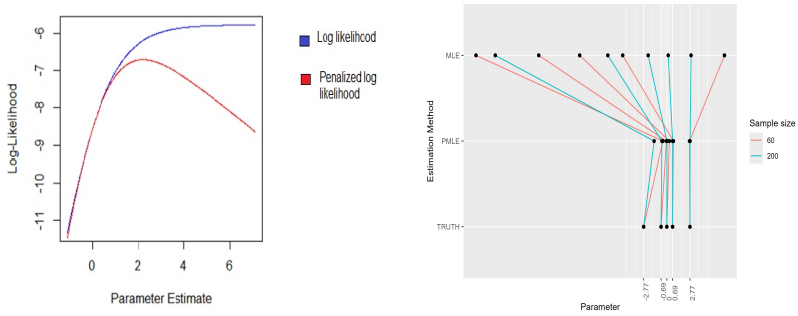
1. **Semiparametric methods like exact permutational method:** In this method, for Poisson regression, the exact permutation distribution of a parameter is obtained conditional on the observed sufficient statistics of all other parameters. The conditioning on the sufficient statistics of nuisance parameters was first used by Cox (1970) for Fisher's exact test for 2×2 contingency table. In such permutation methods, all possible contingency tables are listed down which satisfy a condition of the fixed row sums and the fixed column sums. Conditional inference for a parameter of interest is obtained by conditioning on sufficient statistics of nuisance parameters. Later extensions of this method for higher dimensions were developed (Mehta and Patel, 1995; Mehta et al., 1992; Tritchler, 1984). The maximum conditional likelihood estimate (MCLE) is the value of the parameter that maximises its exact conditional likelihood. If a finite MCLE does not exist, it can be replaced by a median unbiased estimate (MUE) (Hirji et al., 1989). If the exact distribution is degenerate, neither MCLE nor MUE can be computed.

2. **Parametric methods (frequentist):** Firth penalisation of the likelihood is often used for reduction of small sample bias in the maximum likelihood estimation in GLMs. Firth (1993) suggested adding a penalty term to the log-likelihood of the exponential family of distributions, such that the

penalised log-likelihood becomes:

$$l^*(\beta; \mathbf{y}, \mathbf{x}, \mathbf{z}) = l(\beta; \mathbf{y}, \mathbf{x}, \mathbf{z}) + (1/2) \log |I(\beta)| \tag{2.1}$$

where l denotes the log-likelihood under a GLM; β denotes the regression coefficient or parameter of interest including intercept; $|I(\beta)|$ denotes the determinant of the Fisher information matrix; and $\mathbf{y}, \mathbf{x}, \mathbf{z}$ are the vectors of the responses, covariates and the offset variable for n individuals. The Poisson rate is λ_i such that $\log(\lambda_i) = x_i\beta + z_i$. More details about the Poisson model are given in subsection *Computational aspects*. Parameter estimates obtained using the Firth penalisation method are shown by FL in article **I**. They are also known as MPLE or PMLE (penalised maximum likelihood estimate) in some other software or publication. In the case of separation or quasi-separation, the likelihood becomes monotonic. The Firth penalization method reduces the bias and makes the likelihood non-monotonic. This enables obtaining a finite parameter estimate. The effect of Firth penalisation on the monotone likelihood is shown in Figure 2.2(a).



(a) Penalized loglikelihood **(b)** Shrinkage

Figure 2.2: Properties of the Firth penalisation method. (a) Monotone log-likelihood in the case of separation and corresponding penalised log-likelihood. (b) The shrinkage property of penalisation. This graph is generated from the simulation study in Article **I**. Truth shows the actual parameter value, MLE shows its estimate using ML and FL shows its estimate using the penalised likelihood method.

Another advantage of Firth penalization is the shrinkage of parameter estimates ensuring smaller standard errors (Firth, 1993). Parameters shift towards zero as shown in Figure 2.2(b) obtained from the simulation study described in Article **I**.

Computational aspects

The log-likelihood of the regression coefficient β under the Poisson model

is given by

$$l(\beta; \mathbf{y}, \mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} [-\exp(x_i \cdot \beta + z_i) + (x_i \cdot \beta + z_i) y_i - \log(y_i!)],$$

where $i = 1, 2, \ldots, n$ denotes individuals in the data. For the i^{th} individual, y_i is the response that is the number of events, $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is the covariate vector of p covariate values and z_i is the offset term. The response variable y_i is assumed to follow the Poisson distribution. The first covariate is assumed to be 1 for all individuals and the corresponding coefficient or parameter is an intercept term of regression. The offset is generally the logarithm of the follow-up time or the logarithm of the number of observations for the i^{th} individual. When there are multiple observations for the same person, they can be treated as Bernoulli trials with an event indicated by one and a non-event by zero. The Poisson model is a GLM with a log link which means the linear predictor term $(x\beta + z)$ is connected to the Poisson mean by a log link. The ML estimates of the regression coefficient $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ are obtained by solving the score equations

$$\frac{\partial l}{\partial \beta_k} = \sum_{i=1}^{n} [x_{ik} \cdot y_i - x_{ik} \exp(x_i \cdot \beta + z_i)] = 0, \text{ for } k = 1, 2, \ldots, p.$$

The estimates of the Firth penalised likelihood can be obtained in two ways, either by actually adding the penalty term in the likelihood and then maximising it or adding some non-zero term in the response iteratively (Firth, 1993).

- Adding a penalty term to the log-likelihood: In this approach, the penalty term is computed by using the observed information matrix and the penalised log-likelihood is obtained by adding the penalty to the log-likelihood. The estimates are obtained by solving the modified score equations iteratively.

  $$\frac{\partial l^*}{\partial \beta_k} = \frac{\partial l}{\partial \beta_k} + \frac{1}{2|I(\beta)|} \text{diag} \left[\frac{\partial I(\beta)}{\partial \beta_k} \right] = 0 \text{ for } k = 1, 2, \ldots, p$$

  for p coefficients of covariates. The log link is the canonical link for Poisson distribution, so the expected information is the same as the observed information. Computing the observed information matrix and its derivative is quite straightforward for Poisson regression. The

$(r,s)^{th}$ entry of the observed information matrix is

$$\sum_{i=1}^{n}(x_{ir}x_{is}\cdot\exp(x_i\cdot\beta+z_i)),\ 1\le r,s\le p$$

and $(k,r,s)^{th}$ entry of its derivative $\partial I(\beta)/\partial\beta$ matrix is

$$\sum_{i=1}^{n}(x_{ik}x_{ir}x_{is}\cdot\exp(x_i\cdot\beta+z_i)),\ 1\le k,r,s\le p.$$

I, along with my colleagues at Cytel Inc., had developed software to apply this method to many GLMs ("LogXact 12, Cytel Inc. Cambridge, MA. Software Manual", 2019).

- Pseudo-response by using the projection or hat matrix: Score equations can be modified by using a pseudo response instead of the observed response. Firth(1993) showed that, for GLMs, penalisation can be applied by solving the modified score equations iteratively with an additional non-zero term $(\kappa_{3i}h_i/2\phi\kappa_{2i})$ to the response y_i in each iteration, where κ_{ji} denotes j^{th} cumulant, ϕ denotes the dispersion parameter in GLM and h_i is the i^{th} diagonal entry of the hat matrix H or leverage of i^{th} term in the model fit. For Poisson distribution, all cumulants are equal to the Poisson parameter λ and dispersion is one. So, the modified score equation is

$$\frac{\partial l^*}{\partial\beta_k}=\sum_{i=1}^{n}x_{ik}[y_i+(h_i/2)-\exp(x_i\cdot\beta+z_i)]=0\text{ for }k=1,2,\ldots,p,$$

$$H=XW^{1/2}(X^TWX)^{-1}XW^{1/2}\text{ with }W=\text{diag}(\exp(x_i\cdot\beta+z_i)).\quad(2.2)$$

From this form of the modified score equations, it is clear that FL estimates can be obtained by ML estimation on augmented data that consists of the original data in which observed outcomes y_i were augmented by $h_i/2$ in each iteration. We used this Firth estimation method in Article **I**.

3. **Bayesian methods:** The penalty term, added to the log likelihood, as described earlier, resembles the Jeffreys invariant prior. By exponentiating the equation (2.1), it can be seen as the posterior distribution of β (up to the proportionality constant) is equal to the product of the likelihood and the prior,

$$P^*(\beta)\propto L(\beta;\mathbf{y},\mathbf{x},\mathbf{z})\times\sqrt{|I(\beta)|},$$

where P* denotes the posterior distribution of β, the first term $L(\beta; \mathbf{y}, \mathbf{x}, \mathbf{z})$ is the likelihood function of β for the given data and the second term can be interpreted as the prior for β. Note that the Jeffreys prior for the parameter β is proportional to $\sqrt{|I(\beta)|}$ and is a non-informative prior in the sense that it is invariant under reparametrisation.

This shows that the penalised likelihood method of bias reduction has Bayesian interpretation. From the Bayesian perspective, obtaining the maximum penalised likelihood estimates is equivalent to obtaining the posterior mode using Jeffreys non-informative prior (Firth, 1993; Kosmidis and Firth, 2009).
Another choice of prior distribution for β is the normal distribution. The spread of the normal prior for a regression coefficient is determined by the width of a prior interval for the associated risk ratio $\exp(\beta)$ since, relative risk or risk ratio $\exp(\beta)$ is generally more informative for medical practitioners. Hence, using a prior for $\exp(\beta)$ may be more relevant than a prior for β. Instead of specifying the prior explicitly, (Sullivan and Greenland, 2013) showed that prior information can be specified in terms of augmented data. For each parameter of interest, one pseudo- observation is added with a value (an approximation constant) for the associated covariate and 0 for all other covariates. No pseudo-observations are specified for the intercept. This method avoids the problems of computing time and the tedious diagnostics of convergence of Markov chain Monte Carlo algorithms in the Bayesian analysis. Since prior distributions can be represented by pseudo observations, one for each regression coefficient, with a value (an approximation constant) for the associated covariate, and 0 for all other covariates. The maximum posterior estimates can be computed by applying maximum likelihood methods to the augmented data resulting from adding the pseudo observations to the original data set.

2.2 Bias reduction in prediction

In Article **I**, we applied all methods discussed in the earlier section to Covid data and Implant dentistry data. All the methods mentioned above give bias reduced estimates of regression parameters, but they may not give a bias reduced prediction of the response. We studied the predictive properties of all these methods for Poisson regression for count data in Article **I**.
To our knowledge, no available computer software provides an estimate of the intercept term in exact Poisson regression. So, it was not possible to include

this model when evaluating prediction performance of the different modelling approaches. The Bayesian data augmentation method gave an unbiased prediction when priors with large variance were specified as discussed in Article **I**. In the Firth penalisation method, it is generally seen that the removal of bias from regression coefficients of covariates introduces some bias in the intercept term. If the model with bias reduced regression coefficient estimates is used for prediction, then the bias in the intercept induces bias in the prediction, making the predicted number of events higher than the actual number of events. To avoid this problem, we extended two methods that were initially proposed for binary logistic regression (Puhr et al., 2017) - 'FL with intercept correction' (FLIC) and 'FL with added covariate' (FLAC). In the FLIC method, the parameter estimates are obtained using the Firth penalisation method. Then ML is again applied for estimating only the new intercept term while treating the earlier estimated linear predictor term as offset. In the FLAC method, an augmented data set is constructed by adding a pseudo observation for each original observation with the event count corresponding to the penalty $h_i/2$, where h_i indicates the entry of the hat diagonal matrix (2.2). A binary extra covariate, which distinguishes the original from the pseudo-observations, is defined. The maximum likelihood estimation method is used on these augmented data to obtain the FLAC estimates. While using FLAC estimates for prediction, the extra covariate is set to zero. In Article **I**, we extended the use of these methods to Poisson regression. We showed that these two methods, FLIC and FLAC, give exactly the same parameter estimates for all covariates and also for the intercept term in a Poisson regression. Further, both methods give unbiased predictions in Poisson regression. With a simulation study, we showed performance of all these methods in terms of parameter estimation for small, medium and large sample data with varying incidence of separation. Confidence intervals for these estimates are also discussed in Article **I**. In dental implant study, the predicted complications using FL method were 39.5 as compared to the true 37. FLIC and FLAC, both methods, predicted the number of complications correctly as 37. In this thesis, all interval estimation topics are discussed in more detail in section 6.1.

Many other regularisation methods like Lasso and Ridge are used for various improvements, like variable selection, or shrinkage in regression coefficients of multiple regression models. The R package glmnet and the references therein can be referred to for more details. Many authors suggest excluding intercept term while applying penalisation methods (Efron and Hastie, 2016). Some machine learning literature consider the intercept term itself as a bias term (Hastie et al., 2009). Sullivan and Greenland (2013) caution about the use of the intercept term and its interpretability in the model. They discuss methods of penalisation by

using Bayesian prior for regression parameters both with and without intercept. Lasso penalty has a Bayesian alternative with Laplace or double exponential prior and Ridge penalty has a Bayesian alternative with Gaussian prior. The links between the penalisation methods and their Bayesian alternatives are studied in the literature (Mallick and Yi, 2013; Park and Casella, 2008).

Connection between Poisson and binomial setting

All methods we studied in Article **I** were earlier developed for binary logistic regression. We extended their use to Poisson regression for count data. Both binary logistic and Poisson log-linear models are GLMs. Both these regressions use their respective canonical link to the distribution of the response. Extending the methods developed for binary logistic regression to another GLM paved the road for other regression methods in GLM. In both example datasets we used in Article **I** had count data, that is, the number of events as response. However, it can be viewed as Poisson approximation to binomial. In Covid data, since the tests were performed on only one occasion, the presence or the absence of infection for each person at that time made the response binary. The problem of number of infections in each group can be considered as comparison of binomial probabilities of being infected in each group. Also in Implant dentistry data, each patient had undergone one or more dental implants and we counted number of complications as the number of events. This again can be considered as binomial probability of failure or complications of the implant for each individual. In Poisson settings, logarithm of number of implants are taken as offset and number of complications as events. In binomial setting, number of implants can be viewed as number of Bernoulli trials for each patient and complication can be viewed as a binary response of failure with the assumption that each Bernoulli trial has the same probability of complication and Bernoulli trials for the same patient are independent. In Article **I**, we used only one type of event which can be considered as an event related to only one treatment response. However, disease activity or effect of treatment may be shown by more than one response variable or disease marker. Henceforth I will use the term disease marker in connection with the treatment response data. When multiple markers are indicators of the disease activity and they give rise to different types of events, studying the dependency of these markers is necessary. I discuss simultaneous inference from multiple markers in the next chapter.

Chapter 3

Comparisons of groups, effect measures and multiple markers

In this chapter, I discuss comparisons of two groups with respect to multiple markers. If two groups are of patients receiving some treatment for a disease then, disease markers can indicate treatment response. While comparing two groups, the frequently used measures of comparison are difference in treatment responses or ratio of treatment responses, though there exist other measures. Henceforth, for the measure of comparison the term effect measures is used, the name commonly used in clinical studies for measuring the difference in treatment responses in two groups. There are many challenges in getting a combined inference from multiple markers. I list some of them here.

Different markers may be of different data types: Combining the markers of the same data type is much easier by means of joint multivariate distributions. For example, if all the markers are of continuous type, then a multivariate normal distribution is a common choice for the combined analysis. The joint distributions of all markers from two groups can be compared by using median, mode, or percentiles. However, for mixed type of data like wet-AMD real-world data, the markers VA and CFT are of continuous type while the fluid status markers are of binary type. It is difficult to get the joint distribution for such markers. One possibility is to determine a distribution of VA for each category of the fluid status. However, this seemingly simple task gets complicated if the data on all markers are not available at the same time. In multivariate studies, a commonly used approach is the copula method where for each outcome a marginal distribution is specified and a correlation structure among all outcomes is specified (Größer and Okhrin, 2022). When the dependence of multiple variables is not known or when the dependence structure varies with time, copulas may not be useful

(Fermanian and Scaillet, 2004).

Same data type but different effect measures: Even for the same data type, the effect measure can be different for each marker. For example, for one binary marker a commonly used effect measure may be a difference in proportions, while for another binary marker, a ratio or odds ratio of proportions may give a better understanding of the comparison. Getting a common or joint inference from different effect measures from various data-types poses a substantial challenge.

Hierarchy among markers for clinical decision making: While obtaining a common inference from multiple markers, the hierarchy among markers needs to be considered. Not all markers are primary indicators of the disease progression or of the treatment response. In clinical trials, multiple comparison and closed testing techniques of hierarchical hypothesis testing are often used in such scenarios. Their nonparametric alternatives, like nonparametric combination tests, (Pesarin, 1990) are beneficial for small sample studies. For example, in wet-AMD real-world data, the fluid status or CFT is the primary indicator of the treatment response being the treatment target, while changes in VA may occur due to changes in the primary indicator.

Not the same inference from different markers: The treatment may not affect all markers the same way. If one marker suggests a good treatment effect while another marker suggests the opposite, getting the resultant effect becomes a complex task.

In the frequentist approach, hypothesis testing is commonly used for comparing two groups. In clinical trials, for comparing multiple groups or multiple treatments, multiple comparison procedures such as Bonferroni adjustment are used (Hochberg, 1988; Simes, 1986). The Bonferroni correction divides the overall type-I error α by the number of hypotheses m and tests each individual hypothesis at a significance level of α/m. Many variations and improvements of the Bonferroni method are used in practice (Rosenfeld et al., 2006). In such methods, the maximum allowed type-I error is divided by the number of multiple hypotheses which gives a small type-I error for each individual test. This error could be so small that it is almost impossible to get a tail probability smaller than this error rate. This means that it is not possible to reject any hypothesis. This is especially true for small sample data where distributions are discrete and in general it is not possible to utilise all the type-I error. The Bayesian alternative of multiple

hypotheses testing is also employed quite often (Abramovich and Angelini, 2006; Guo and Heitjan, 2010; Scott and Berger, 2006). Another popular method is the nonparametric combination test especially suitable for small sample data. In these methods, multiple markers are assessed with multiple effect measures, the marginal p-values are obtained as the first step and then the combined inference is carried out by a combination or an adjustment of marginal p-values (Pesarin, 1990; Westfall and Krishen, 2001). Most of these procedures rely on p-values of the individual marker. In recent years, permitting the whole inference to rely on a single number, p-value, has come under criticism (Sheldon, 2019; Wasserstein and Lazar, 2016). Instead of combining individual inferences for each marker or hypothesis, we proposed a Bayesian method of obtaining a joint posterior distribution for all markers and rely on the inference on simultaneous credible intervals in Article **II**. Therein, our aim was to compare two groups by using the simultaneous inference from multiple markers. In the wet-AMD case-study, we considered two groups based on baseline age of patients with wet-AMD. Since wet-AMD is an age-related disease, we wanted to study how the disease markers differ in the older and younger patients' groups. In another example, we compared the markers at the baseline and the markers after six months of treatment. In either of the examples, our interest was not in the selection or rejection of a treatment but to compare two groups with respect to multiple markers simultaneously; hence, instead of hypothesis testing, we used simultaneous credible regions of all markers. Studying the joint posterior distribution of multiple markers in each group and comparing these distributions help in understanding how the two groups are similar or different with respect to those markers. From these joint posterior distributions, it is possible to examine posterior marginal distribution or joint distributions of a subgroup of markers to divide markers into groups with similar effect. This helps in understanding the correlation among markers. Unlike multiple comparisons, we do not divide the required type-I error. Instead we use a required credible level to get simultaneous credible regions of multiple markers.

As already seen in Chapter 2, the effect of penalisation can be achieved by using the Bayesian method with a non-informative or a weakly informative prior. Further, the shrinkage property gives shorter credible intervals. In Article **II**, we use weakly informative priors to achieve this effect. Another advantage of the Bayesian method is the possibility of incorporating earlier knowledge or history in terms of the prior or augmented data. In the next section, the method of obtaining simultaneous credible regions for multiple binary markers is described in brief.

3.1 Contingency tables and simultaneous credible intervals

In Article **II**, We considered two or more binary markers and our interest was in their simultaneous or combined effect. For this purpose, contingency tables provide an appropriate framework. The method for a single binary marker discussed in Nurminen and Mutanen, (1987) showed an excellent way of obtaining the inference for different effect measures for the same marker. This method uses prior information of the natural parameters of the marker instead of prior information for each effect measure separately. If the disease has already been studied earlier in observational or randomised studies, then there could be information available for the natural parameters of its markers, for example, in wet-AMD disease the probability of the presence (or absence) of intra-retinal fluid (IRF) is commonly reported. Since not all previous studies may have used the same effect measure for the same marker, it is difficult to get prior distribution for each effect measure using all the previous information. To utilise all the previous information about a particular marker, it is necessary to consider the prior distribution on the natural parameters of the marker and not for each effect measure separately. For example, when comparing two groups with respect to the presence or absence of IRF in case of wet-AMD, the effect measures could be the difference in the probabilities of the presence of IRF in the two groups or their ratios or odds ratios. Our approach is to specify prior for the probabilities themselves and then provide the posterior summary of the difference or the ratio. We extended this method to multiple binary markers in Article **II** to get a simultaneous inference of different types of effect measures from multiple binary markers. We used Migraine clinical trial data and wet-AMD real-world data for illustrating the use of simultaneous credible regions. Migraine clinical trial data had all the binary markers while in wet-AMD real-world data fluid related markers were binary, while VA and CFT were continuous markers. While comparing the treatment effect at six months in two groups of patients, we dichotomised VA based on the comparison to the baseline value, that is, VA above the baseline was marked as one and VA below the baseline was marked as zero, only for illustration purposes, though such comparisons have been used in the wet-AMD context. For a single binary marker, there are two possible outcomes 0 and 1, for two binary markers there are four possible outcomes viz., $(0,0),(0,1),(1,0),(1,1)$ and for m binary markers there are 2^m possible outcomes. In this case the natural parameters are the probabilities of these 2^m possibilities. We used the Dirichlet prior obtained from either the historical data or some weak-informative priors for these probabilities, and the

multinomial distribution was used as the data distribution to compute the joint posterior distribution for all parameters of all markers (Mosimann, 1962). From the joint posterior distribution, it was possible to obtain the marginal posterior distribution for each marker, and more importantly the marginal distribution for any specific effect measure for each marker by drawing a sufficiently large number of samples from the posterior distribution. We used the Jeffreys prior for parameter estimation in GLM extensively in Article **I**. In Article **II**, we extended the use of the Jeffreys prior to multiple markers. To analyse the contingency tables, GLMs are used for a long time (Nelder, 1974). Jeffreys prior for a binomial probability is proportional to $\beta(1/2, 1/2)$. Adding a half to each cell of a contingency table is also known as the continuity correction, and is used especially in sparse contingency tables, since this technique avoids zero cell counts and enables estimation of the cell probabilities. However, for multiple markers, say for m markers, there are 2^m cells in one contingency table. So, in order to apply Jeffreys prior to a contingency table of m markers, one has to add 2^m times $1/2$ to the sample size, which is not desirable for small sample studies with large m. For multiple markers, as suggested in Berger et al. (2015), we applied an objective prior, also known as a reference distance prior, which adds only one to the whole sample size. The reference distance prior for a contingency table with 2^m cells is a Dirichlet prior $D(\alpha)$ where each parameter is $\alpha = 1/2^m$. We compared posterior distributions obtained using the uniform prior, the Jeffreys prior, and a reference distance prior for multiple markers.

After the posterior distribution for multiple binary markers was obtained, we proposed algorithms to compute posterior credible regions (or intervals) for parameters (or effect measures) of multiple markers simultaneously in Article **II**. We discussed two types of credible regions commonly used in Bayesian inference: equal tailed posterior credible regions (or intervals) (ETI) and highest posterior density regions (HDI) (Kruschke, 2014).

Equal tailed credible region

The simultaneous equal tailed credible region (ETI) for multiple parameters is the joint posterior region obtained such that its corresponding marginal posterior intervals are equal tailed two-sided credible regions for each parameter with equal tail probabilities for all parameters. This region was obtained by applying a bisection method on the tail probability for an individual parameter.

Highest posterior density set

A Highest posterior Density Interval or region (HDI) may not be a single connected region. It is defined such that the density at any point inside the HDI is more than the density at every point outside the set. The HDI is computed by ordering the point densities and finding a threshold k such that all the points with

their ϵ neighbourhood density more than or equal to k are included in the HDI and all the points with their ϵ neighbourhood density less than k are excluded. A more detailed algorithm is given in Article **II**. In Article **I**, we had discussed the wald type and profile likelihood based frequentist intervals. Similarities and difference of the Bayesian intervals ETI and HDI with these frequentist intervals are discussed in section 6.1.

These simultaneous credible regions helped in getting higher density regions of the parameter space of multiple markers. Like many methods, this method also had limitations when used at higher dimensions. Visualising simultaneous credible regions of high dimensional data was not practical. To overcome this issue, we demonstrated the use of joint cumulative probabilities of a subgroup of markers. This method helped in identifying subgroups of markers with a similar effect. Such subgroups of markers suggested that we can use this method for gatekeeping procedures used in a clinical trial for hierarchy of hypotheses.

In clinical trials, when multiple markers are tested together and if multiple markers have a natural hierarchy of importance based on the treatment effect, then the markers are grouped into families according to their importance. These families are referred to as primary family, secondary family, in the order of their importance. These families can be a family of markers or a family of hypotheses to be tested in a particular order. Gatekeeping procedures are applied to these families to address the multiplicity problems by explicitly taking into account the hierarchical structure (Dmitrienko and Tamhane, 2007). I will review two main types of gatekeeping procedures here.

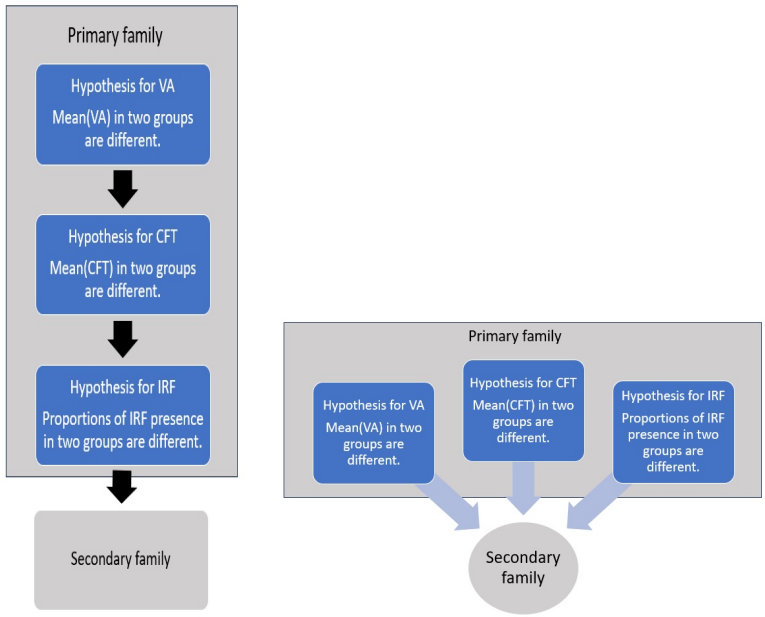
3.1.1 Gatekeeping

In gatekeeping procedures, each higher family of markers plays the role of a gatekeeper for the next family. Two types of gatekeeping procedures are often used. They are illustrated in Figure 3.1 with an example of three wet-AMD disease markers in the primary family: VA, CFT, and IRF. It is assumed to have one hypothesis each for each marker to decide equality in two subgroups. Let us assume that all other markers are in the secondary family.

1. Serial gatekeeping: In this method, all markers (hypotheses) are tested serially. If the available data support the hypothesis that each marker from the higher ranked family shows a non-zero treatment effect, then the lower ranked families will be assessed as shown in Figure 3.1 (a).

2. Parallel gatekeeping: In this method, the markers from the same family are tested in parallel. If the available data support the hypothesis that there is a non-zero treatment effect for at least one marker from the higher ranked family,

then the lower ranked families will be assessed as shown in Figure 3.1 (b).
A more complex decision on hierarchical testing can be made by using tree-structured gatekeepers or graph based gatekeepers (Röhmel et al., 2006).



(a) Serial gatekeeping **(b)** Parallel gatekeeping

Figure 3.1: Flow-chart illustrating serial (a) and parallel (b) gatekeeping procedures with a simple example of a primary family of three markers of AMD: VA, CFT, and IRF. There is one hypothesis each for each marker to decide equality in two subgroups.

We illustrated the use of the simultaneous credible region for gatekeeping in Figure 8 of Article **II** by using Migraine clinical trial data. The equal tailed simultaneous credible region and marginal credible regions corresponding to two markers, pain freedom at 2 hour post treatment EP_1 and sustained pain freedom up to a 24 hour period EP_4, from Migraine clinical trial data are shown by shaded regions. These regions are computed at 90% credible level for the effect measure *difference of proportions*. The grey area in the serial gatekeeping figure shows the joint posterior region and each stripe in the parallel gatekeeping figure shows the marginal posterior region for one marker with no restriction on the other marker. The figure can be interpreted on the lines of hypothesis testing but focusing on the acceptance rather than the rejection of a hypothesis or the regions that support the hypothesis. A higher difference of proportion indicates a higher treatment effect. If the grey area of the serial gatekeeping lies in the upper right region of the point of the sufficient treatment effect, say $(p_1 > 0, p_2 > 0)$, then all the

markers show the required non-zero treatment effect. It indicates that the family of markers has passed through the serial gatekeeping. In parallel gatekeeping, if the point of the sufficient-treatment effect is less than the lower limit of one of the grey stripes but not less than the lower limit of other grey stripe, then there is at least one marker showing a non-zero treatment effect but not the other. It indicates that the family of markers can pass through the parallel gatekeeping. Based on the context, the reference point of sufficient treatment effect and the simultaneous credible region can be compared to obtain the gatekeeping inference.

In Article **II**, we developed the methods for obtaining a simultaneous credible region of multiple binary markers where the Bayesian posterior credible intervals were used instead of p-values for gatekeeping. Since the sample sizes in the groups to be compared were known and fixed in our dataset, and at least two of the markers were binary, we developed the methods for multiple binary markers. A 2×2 contingency table is used while testing a single marker in two groups. We illustrated multiple markers by using higher dimensional contingency tables, and used the Dirichlet prior for cell probabilities. If the group sizes are not known or the data are not convertible to the binary format, similar methods can be developed by appropriately choosing the data distribution and priors. For example, the Poisson likelihood for counts is appropriate in such situations. The gamma distribution is the conjugate prior for this likelihood. Visualization of simultaneous credible regions for more than two dimensions was not trivial, hence for higher dimensional posterior distribution, we plotted the cumulative joint probabilities.

3.2 Longitudinal measurements of multiple markers

In chronic diseases, patients are followed up for a longer period and they are often exposed to the treatment for an extended period. Disease markers are measured for each patient multiple times during the follow-up. Such data involving repeated measurements of the same variable over a long period of time are referred to as longitudinal data. The timings of the measurements may be regularly spaced or irregularly spaced and may also vary among the patients. If all patients have repeated measurements with the same frequency, the data are considered as balanced data. If all the patients do not have repeated measurements with the same frequency, it is considered as unbalanced data. Analysis of multiple markers or outcomes measured or observed repeatedly over time is of interest in clinical studies. For example, among patients in the wet-AMD study, how the changes

in IRF are reflected in the changes in visual acuity can be an important research question. The measurements of multiple markers for all patients at a fixed time are called cross-sectional data. The methods discussed in the earlier sections and chapters, which deal with a single or multiple markers at a fixed time point, are suitable for such cross-sectional data (Smith, 2015). Methods developed in Article **II** for simultaneously analysing multiple markers can also be used for multiple time points as shown in one of the examples where the treatment effect was obtained in the first six months of the wet-AMD study by comparing the baseline and six month measurements.

Longitudinal data on multiple markers of a disease are collected to study the disease activity, as characterised by the changes in the markers and in their association over time. For a good review of methods on longitudinal data on multiple markers, I refer to Verbeke et al. (2014). The timing between the measurements depends on many factors, such as the rates of fluctuations in the markers, the ease of measurements, and the cost of measurements. For multiple binary markers data at multiple time points, repeated contingency tables can be studied by extending the methods discussed in Article **II**.

3.2.1 Extension to time-dependent contingency tables

For the wet-AMD data, half yearly measurements can be of interest to decide the treatment regime for the next six months. The contingency tables discussed in the previous section can be analysed at $0, 6, 12, 18, 24$ months separately at first. Then, to understand the time-dependency, all contingency tables can be analysed together. Some methods like GEE and log-linear models for independent markers or markers with known correlation have been applied by researchers with the help of existing statistical software (Ziemer et al., 2018). From the Bayesian perspective, a Dirichlet process prior can help in modelling multiple contingency tables and time-dependent correlation between them (Gelman et al., 2003; Ghosal et al., 2010).

The difficulty in the contingency table based approach is that not all markers are measured at all time points, which results in missing data. For example, in the wet-AMD case, it is very likely that VA is measured at each monthly visit, but IRF are measured at baseline, and then only at intermittent visits like 0 month, 12 months, 24 months. Longitudinal data are difficult to analyse when the missing data pattern is not monotone, and the analysis requires probability models which are hard to verify. The results are sensitive to the model assumptions (Little, 2024; van Buuren, 2018).

When multiple discrete markers are measured at discrete time points, another

approach seen in the literature is multistate models. Multiple states of disease progression are defined by using a combination of categories of multiple markers (Lovblom et al., 2024). This method is useful if all markers are measured at the same time and if the total number of markers and possible states are not large.

The above mentioned methods are not useful when some markers are continuous and cannot be categorised into finite categories. In such case, longitudinal analysis can be used. When the data with repeated measurements of independent covariates and treatment responses are observed regularly over a follow-up period, the methods of longitudinal regression analysis are used in order to understand the relationship between the time-dependent or the time-independent covariates and the time-dependent response. The Bayesian alternative methods are available and are especially useful for a small sample or unbalanced data (Gelman et al., 2003). Nonparametric methods such as splines and penalised splines are used quite often when the distributional assumptions of the classical methods are not met (Aerts et al., 2002).

3.2.2 Clustering of trajectories

In a longitudinal study, understanding the response pattern over time can help in deciding the future treatment. When multiple markers together define the disease status, it is difficult to mark a good or a bad response. Mixed effect models are commonly used in the scenarios of clustering with multiple markers (Fitzmaurice et al., 2011; Gelman and Hill, 2006; Mukhopadhyay and Gelfand, 1997). Many R packages with frequentist and Bayesian approaches are available for clustering the individuals based on their response trajectories. R packages, like lcmm (Proust-Lima, 2024)(frequentist approach of latent class models) and mixAk (Komárek, 2024) (Bayesian approach with mixed growth model), to name a few, deal with the clustering problem of multivariate responses of continuous, binary, count, or ordinal type. To our knowledge, all the methods handling the clustering assume a fixed membership of the latent class or of the underlying cluster throughout the study period. These methods are useful for getting an overall idea of how the treatment works for a patient if there are no major changes in the treatment response over the period. If the response trajectories are not linear over time, some packages have also used non-linear trajectories by using the spline interpolation (Dziak et al., 2015; Song and Lu, 2010). For the comparison of various frequentist methods of clustering, I refer to a review article by Den Teuling et al. (2023). In this article, five methods of clustering are compared by their performance on simulated data. The methods are namely-1. Longitudinal K-means, 2. A two-step clustering approach of modelling the trajectories

using a growth curve model (GCM), and clustering the the random effects using kmeans, 3. A group-based trajectory model (GBTM) creates clusters of longitudinal dataset in terms of a mixture of group trajectories, without regard of within-group variability, 4. Growth mixture model (GMM): a method for identifying heterogeneous subgroups in the data using a mixture of growth curve models, 5. In a time-varying effect model (TVEM), the regression coefficients that describe the relation between covariates and outcome can vary over time. The performance was evaluated on two aspects, first the agreement of the group assignment to the simulated reference and the trend estimation error, as measured by a weighted minimum of the mean squared error (WMMSE). The article concludes that the GMM was found to achieve the best overall performance, followed closely by a two-step approach using growth curve modelling and k-means (GCKM), the latter is preferred for large datasets for its computational efficiency. Longitudinal k-means (KML) and group-based trajectory modelling were found to have practically identical solutions in the case that the group trajectory model of the latter method is correctly specified. Both methods performed less than GMM and GCKM in most settings. The TVEM method applied using the software MixTVEM performed poorly on all aspects. For the Bayesian methods of clustering, I refer to the review article by Wade (2023). This article discusses the model-based approach and the loss-based approach of Bayesian clustering along with a discussion on the importance of the kernel or loss selected and the prior specification. The author argues that much of the debate and asymptotic theory focuses on the marginal posterior of the number of clusters; however, quite a different inference is obtained when estimating the full clustering structure.
To explore the usefulness of clustering in analysing the small wet-AMD data wet-AMD real-world data I used R package mixAK . The package fits the Mixture multivariate generalized linear mixed model (MMGLMM) for the multivariate longitudinal outcome by running MCMC for a generalized linear mixed model with several response variables and normal mixtures in the distributions of random effects. Here, a short description of a model is given, avoiding excess technical details, that was used for exploratory analysis only. Three response variables were modelled with distributions according to their datatypes - VA (Gaussian with identity link), CFT (Gaussian with identity link)) and IRF (binomial with logit link). GLM was fitted for each one of them with the linear predictor term including an intercept and a slope term for the time variable.
Notations:

$$Y = (Y_1, Y_2, Y_3) : \text{ three markers (VA, CFT, proportion of IRF presence) ,}$$

N : number of patients or individuals,

n_i : total number of visits of the i^{th} patient,

$\alpha = \{\alpha_1\}$: a vector of fixed effects,

$\beta_i = (\beta_{i,1}, \beta_{i,2}, \ldots, \beta_{i,5})$: a vector of random effects for the i^{th} patient given the cluster membership U_i,

$\eta_{(r)}$: a chosen link function for response $r = 1, 2, 3$,

U_i : latent class membership of patient i,

K : total number of clusters, (K = 2 in the example below),

t : time of the clinical visit.

Model specification:

$$\eta_{(1)}\{E(Y_{1,i,j}|\beta_{i,1}, \beta_{i,2})\} = \beta_{i,1} + \beta_{i,2} \cdot t_{i,j},$$
$$\eta_{(2)}\{E(Y_{2,i,j}|\beta_{i,3}, \beta_{i,4})\} = \beta_{i,3} + \beta_{i,4} \cdot t_{i,j},$$
$$\eta_{(3)}\{E(Y_{3,i,j}|\alpha_1, \beta_{i,5})\} = \alpha_1 + \beta_{i,5} \cdot t_{i,j},$$
$$\text{for } i = 1, 2, \ldots, N \text{ and } j = 1, 2, \ldots, n_i.$$

Other assumptions of the model are as follows.

- For each $i = 1, 2, \ldots, N$, the conditional distribution of the joint random effects vector β_i given the cluster allocation $U_i = k$ (patient i belongs to cluster k), is a multivariate normal with unknown cluster-specific mean and unknown cluster-specific positive definite covariance matrix. The cluster-specific indices are suppressed for conciseness. All unknown model parameters including alpha, cluster-specific means, variance-covariance matrix are collectively shown as vector θ henceforth.

- For each $i = 1, 2, \ldots, N$, the random variables $Y_{i,.}$, are conditionally independent given the random effects vector β_i.

- Random effects vectors $\beta_1, \beta_2, \ldots, \beta_N$ are independent.

The package used a Bayesian approach based on the output from the Markov chain Monte Carlo (MCMC) simulations to infer the unknown model parameter vector θ and to perform clustering. The latent quantities (random effects β_i and component allocations $U_i; i = 1, 2, \ldots, N$) were considered as additional model

parameters with the joint prior distribution for all model parameters specified hierarchically. An MCMC sample S_M of M records was generated from the joint posterior distribution $p(\theta, \beta_1, \beta_2, \ldots, \beta_N, U_1, U_2, \ldots, U_N|Y)$. The estimate of the probability of the patient i belonging to cluster k was obtained from the posterior sample S_M as the posterior mean.

$$
\begin{aligned}
p_{i,k}(\theta) &= P(U_i = k|Y, \theta) \\
\widehat{p}_{i,k}(\theta) &= E[p_{i,k}(\theta)|Y] = \int p_{i,k}(\theta)\ p(\theta|Y)\ d\theta \\
&\simeq \frac{1}{M} \sum_{m=1}^{M} p_{i,k}(\theta^{(m)}).
\end{aligned}
$$

In addition, uncertainty in the classification was explored using the highest posterior density credible intervals (HDI).

From the data, two groups or clusters of patients were of interest: the good responders to the treatment, and the bad responders to the treatment. The clustering result for two clusters is given in Figure 3.2.

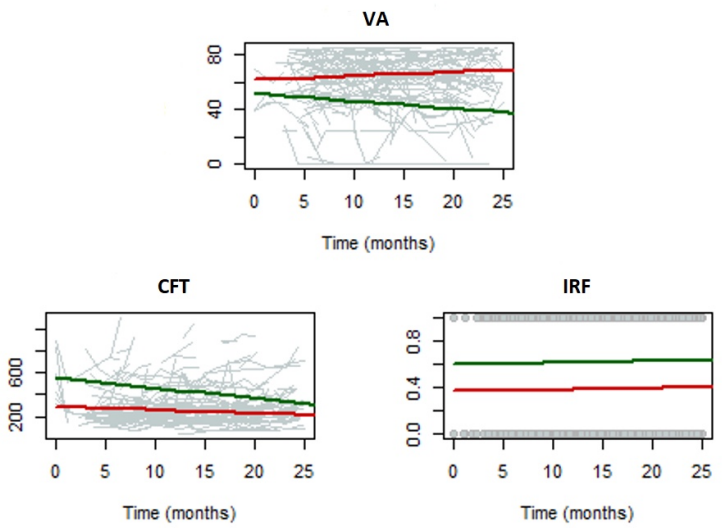


Figure 3.2: Longitudinal trajectories for three markers of the wet-AMD disease: VA, CFT and IRF. 110 patients from wet-AMD real-world data are divided into two clusters using R package MixAK using data from 2 years follow-up. Green and red lines are the mean trajectories for the two clusters.

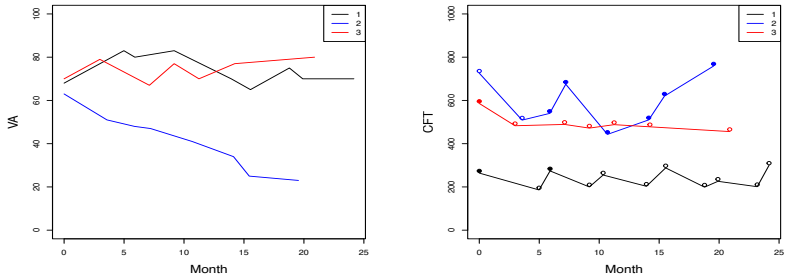
The package MixAK divided the patients into two clusters of sizes 78 and 14 while 18 patients remained unclassified. The red lines in Figure 3.2 indicate,

rightly so, that the patients with a higher VA throughout the follow-up were those with a lower probability of IRF, and with CFT within its healthy range. Let us call this cluster as the cluster of good responders to the treatment. The green lines in the figure indicate that the patients with a decreasing VA over the follow-up period were with a higher probability of the presence of IRF and with a very high CFT at the start. This cluster can be considered as cluster of bad responders to the treatment. However, the number of individuals in the cluster of bad responders was too small. So, the number of patients who could not be placed in any of the clusters was more than the size of the bad responders' cluster making the clustering unusable.
Table 3.1 shows posterior probabilities of cluster allocation for three patients; one each from the cluster of good responders, the cluster of bad responder, and the unclassified group. Their VA, CFT and IRF graphs are shown in Figure 3.3.

Table 3.1: Mean and HDI of posterior probabilities of cluster allocation (upto 2 decimal places) for three patients. Patient 1, 2 and 3 were classified into cluster 1, cluster 2, and the unclassified group, respectively.

Patient	Posterior probability of belonging to Cluster 1 (Good responders)		Posterior probability of belonging to Cluster 2 (Bad responders)	
	Mean	HDI	Mean	HDI
1	**0.99**	(0.99, 1.0)	0.01	(0.00, 0.01)
2	0.00	(0.00, 0.06)	**0.99**	(0.94, 1.0)
3	0.43	(0.099, 0.99)	0.56	(0.01, 0.90)

Patient 1 has a higher posterior probability of belonging to the cluster of good responders. From trajectories of patient 1 it can be seen that VA is high and CFT is within the so called healthy range throughout the follow-up. Also, IRF is absent after a few initial visits. Patient 2 has higher posterior probability of belonging to the cluster of bad responders. For patient 2, VA is high initially but then shows a decreasing trend, CFT is varying but much higher than the healthy range throughout the follow-up, and IRF appears after the initial visit and never disappears. Patient 3 is not classified into any of the two clusters, since their mean posterior probabilities of belonging to cluster 1 and belonging to cluster 2, both are near 50%. For patient 3, VA is high but CFT is also high throughout the follow-up, and IRF is present at the baseline but disappears later. Hence, for patient 3, high VA and IRF absent status suggest low disease activity or good treatment effect, but high CFT suggests the opposite, making the classification difficult. It also seems that such clustering methods, though they allow unbalanced data, they may still be sensitive to the number of observations per patient and the spacing between them. These methods were not able to classify patients



(a) VA **(b)** CFT and IRF

Figure 3.3: VA and CFT trajectories of three patients; one each from clusters of good responders, bad responders, and the unclassified group. (a) VA measurements of three patients joined by a line. (b) CFT measurements of three patients joined by a line. If IRF status is available for a visit, it is shown by a circle above the trajectory; the filled circle indicates the presence of IRF and the empty circle indicates the absence of IRF. In all 3 subplots, 1 shows the patient from the cluster of good responders, 2 shows the patient from the cluster of bad responders and 3 shows the patient not classified as a good or a bad responder.

with contradictory responses from multiple markers like patient 3 in the above example. For some patients, the multiple markers were consistent but the overall response changed after the first year of treatment. For changing the allocation of patients to a different cluster after the first year or periodically needs time-dependent clustering.

Time-dependent clustering

None of the packages provided explicit time-dependent clustering. If a time-dependent clustering is of interest, it is not a simple task to examine whether a patient changes the clusters at some time during the follow-up, unless one executes the clustering method multiple times with the data from different follow-up times separately. The posterior distribution of the clustering obtained in the initial follow-up can be used as the prior for the later follow-up to incorporate the information from the initial follow-up for the inference for the later follow-up. But if patients change their cluster over time and when clustering itself is not stable for small sample studies as we saw earlier, it is difficult to perform time-dependent clustering.

A typical probabilistic model is generally a finite mixture model where each cluster has cluster-specific parameters and some weight is assigned to each cluster. The density function of the mixture model is defined as the weighted sum of densities for the data given the cluster-specific model with cluster-specific parame-

ters. Some authors suggest using a Hidden Markov Model (HMM) for modelling the latent class or cluster (Smyth, 1996). In this case, the cluster-specific parameters include transition matrices, observation density parameters, and initial state probabilities. HMMs can be used for time-dependent clustering where observations belong to hidden states (latent clusters) probabilistically rather than permanently. The HMM allows an observation to transition between hidden states based on a probabilistic transition matrix, accommodating dynamic, time-dependent behaviour.

In general, clustering patients according to some markers is useful for exploratory analysis to understand the pattern or combinations of the markers that come together but they are not useful, for example, when the interest is in the choice of a specific treatment for a patient. Further, in chronic diseases, not all patients are followed-up for the same time. Patients may discontinue the treatment or may get cured from the disease. Clustering patterns learned from data of patients with lesser follow-up may not help in deciding the cluster for longer follow-ups. Overall, though clustering helps in identifying patterns in longitudinal multivariate responses, they may not always be useful in personalised treatment decisions. In chronic diseases, often the longitudinal data are collected along with the time to an event of interest. For example, the event of interest can be death, discontinuation of the treatment, or cure from the disease. In such situations, joint modelling of the longitudinal markers and the time to the event is of main interest. This aspect is discussed in the next chapter.

Chapter 4

Time-to-event and associated covariate processes

In this chapter, I discuss statistical methods for modelling longitudinal observations and an event of interest related to the disease. The event of interest can be death, some impairment due to the disease, cure from the disease, time of stopping the treatment due to insufficient response, or any other event related to the disease. When the time to the event of interest is also to be modelled, along with the longitudinal data, then understanding the effect of the longitudinal trajectories on the event of interest is important. For this purpose, the repeated measurements and the survival time are both modelled simultaneously using joint modelling. In the joint model, there are two submodels. One submodel is a longitudinal model, where the longitudinal response trajectories are modelled over covariates. The other submodel is a survival model, where the event of interest is modelled on the covariates. Both the submodels are connected by some common covariates and with the assumption that longitudinal and survival outcomes are independent given the effects of common covariates. The success of this joint modelling relies on correctly specifying the longitudinal model, the survival model and the link between them. A standard joint model involving multivariate longitudinal outcomes and an event of interest consists of a multivariate generalized linear mixed-effects model (MGLMM) for the longitudinal submodel with multivariate longitudinal outcomes and a survival submodel for a time-to-event outcome (Lovblom et al., 2024). A relative risk model is commonly used to model the hazard function in the survival submodel. A relative risk model consists of two parts; a baseline hazard and an exponential term involving longitudinal covariates. The effect of longitudinal covariates can be accommodated by using the latest value of the covariate or slope at time t or area under the trajectory until time t or the random effects parametrization used in the longitudinal sub-model.

The baseline hazard function is generally modelled as a piecewise-constant or a spline function, or parametric specifications, such as the Weibull distribution.
The early development of joint models was motivated by the HIV/AIDS clinical trials (De Gruttola and Tu, 1994; Faucett and Thomas, 1996; Tsiatis et al., 1995); however some time later, joint models became popular with cancer studies, and especially with the quality-of-life data in cancer studies (Chi and Ibrahim, 2005; Kirkwood et al., 2000). For more recent developments in joint modelling and related software, refer to the review article Wang and Zhong (2025). In our datasets of wet-AMD real-world data and wet-AMD clinical trial data, not all markers were measured at the same time. The correlation among anatomical markers IRF, SRF, and CFT and the physical marker VA is not yet clearly known. All these markers together explain the disease activity. They all are disease markers and it is difficult to divide them into a typical regression setting of independent covariates and responses. When the correlation between various disease markers is not clearly known, or when other independent covariates are not available, or when multiple disease markers are not observed together, it is difficult to model any of the submodels in the joint model. Hence, we decided to not use joint modelling for our dataset.
Another approach to model an event of interest and longitudinal covariates is landmark modelling (van Houwelingen, 2011). In landmark modelling, some predecided time points are considered as landmarks. Survival analysis or prediction is performed at each landmark point based on the information available at that landmark time point. Compared to joint modelling, landmark models are less complex and more flexible in model assumptions. Generally, a landmark model is used for prediction of future value using the latest marker value or latest slope of the trajectory. The use of only the latest value ignores the complete history of longitudinal measurements.
To incorporate the information from the whole trajectory, some summary measure is required. If frequent measurements of disease markers are available, the study of the area under the longitudinal trajectory of these markers is a popular method among medical practitioners (Gross et al., 2015; Wilding et al., 2012). When markers are indicative of the disease activity, they are considered as responses to the treatment. Here we refer to them as covariate processes. The area under the longitudinal covariate trajectory can be viewed as an integrated covariate process. The events defined by the integrated covariate process, such as when it reaches a threshold within a certain time can be informative of the disease activity or the treatment effect. To understand the association between these two types of events, one based on the integrated covariate process and another the event of interest, it is required to study the two processes leading to the two

events together. Modelling the trajectory itself may not be necessary in this type of problem. Here, the joint survival probability of both, the type of event times, or the survival function of one type of event conditional on another type of event is useful for decision making. In Article **III**, we proposed estimation methods for such joint survival functions. Before going to the joint survival of the event of interest with the integrated process, here is a review of some nonparametric methods in survival analysis.

4.1 Nonparametric estimation

The main aim of survival analysis is to estimate the survival probability or its functions at any time during the follow-up. The data used in survival analysis contains the observation times which can be the event time or censoring time, and a binary indicator telling whether the event of interest has been observed. In medical studies, the patients are expected to be followed-up until the time of the event of interest. However, in practice, the follow-up may end before that due to reasons related or unrelated to the disease, resulting in censored observations.
If there is no censoring then the probability of survival up to time t can be estimated using the empirical probability, that is, the ratio of those surviving at time t to the total number of individuals in the study. This empirical estimator is a naive unbiased estimator.

4.1.1 Statistical formulation

Let τ be the time to an event of interest and let C be the censoring time. Let the observed time be $T = \tau \wedge C$ and the event indicator be $\delta = \mathbf{1}(\tau \leq C)$. Further, we assume that the censoring is independent of τ. Let the survival function be $S(t) = P(\tau > t)$ and the censoring survival function be $G(t) = P(C > t)$. Let observations from n independent individuals be $\{T_i = \tau_i \wedge C_i, \delta_i = \mathbf{1}(\tau_i \leq C_i),\ i = 1, \ldots, n\}$.

As mentioned above, in the absence of censoring, a survival estimator can be obtained using an empirical estimator as

$$\widehat{S}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(T_i > t).$$

In the presence of independent right censoring, the survival probability has to be updated at each event time. The Kaplan-Meier (KM) estimator of the survival function is a widely used nonparametric maximum likelihood estimator

of the survival function in the presence of independent censoring (Andersen et al., 1992; Kalbfleisch and Prentice, 2002), and is given as follows.

$$\widehat{S}(t) = \prod_{s \leq t} (1 - \Delta \widehat{\Lambda}_n(s))$$

where

$$\widehat{\Lambda}_n(t) = \sum_{i=1}^{n} \frac{\mathbf{1}(T_i \leq t)\delta_i}{\sum_{j=1}^{n} \mathbf{1}(T_j \geq T_i)}$$

is the Nelson-Aalen estimator of the cumulative hazard of τ. Note that the numerator is one for i if the event is observed before t and the denominator is the size of the risk set at the observed event time.

Another method of incorporating the censoring in the estimator is modifying the empirical estimator using weights with the inverse probability of censoring when the event of interest is observed, i.e. $\delta_i = 1$. This method is known as the inverse probability of censoring weighting. The inverse-probability-of-censoring-weighted (IPCW) estimator was first proposed by Robins and Rotnitzky(1992) as an extension of the empirical estimator for the estimation of the survival function in the presence of dependent censoring. The IPCW method can also be viewed as an empirical estimator with a weight assigned to each term, where the weight equals the inverse of the probability of remaining uncensored. Though the IPCW estimator was initially developed for a dependent censoring scenario, it was later used for other purposes, such as for handling missing data (Seaman and White, 2013) or for reducing the bias in the survival estimation during treatment switching. An R package trtswitch (Lu, 2025) includes implementation of the IPCW method for treatment switching. Many authors have used IPCW method since its introduction for many purposes (Cuparić, 2021; Datta et al., 2010; Dewan et al., 2004). A step-by-step implementation of the IPCW method for R users is given in (Willems et al., 2018).

When the censoring is independent of the event of interest then the Kaplan-Meier estimator of the survival function coincides with the inverse probability of the censoring weighted estimator suggested by Satten and Datta (2001) as below.

$$\widehat{S}(t) = 1 - \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\mathbf{1}(T_i \leq t)\delta_i}{\widehat{G}(T_i-)} \right]. \tag{4.1}$$

where $\widehat{G}(.)$ is the Kaplan-Meier estimator obtained by reversing the roles of the event time and censoring time. Another result from Satten and Datta (2001)

showed that when the censoring is independent of the event of interest, then the Kaplan-Meier estimator of the survival function can be written as

$$\widehat{S}(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbf{1}(T_i > t)}{\widehat{G}(t)}\right]. \tag{4.2}$$

It is easy to see that $P(\tau > t) = P(T > t)/P(C > t)$ and replacing the numerator by its empirical estimator gives an unbiased estimator of $P(\tau > t)$. When the censoring survival is to be estimated at some fixed t, one can either use some known censoring distribution, if it is available, or one can estimate the censoring survival function by using the data under analysis with the censoring observed until time t as events. In practice, the unknown censoring survival function is replaced with the KM estimator.

This form (4.2) of the survival estimator at time t puts the risk set at time t in the numerator and the censoring survival estimate at time t in the denominator. It can be viewed geometrically as the whole sample space is divided into two parts - observations after t are used for getting the risk set in the numerator and observations until t are used as the observed information until t from which the censoring events are obtained to estimate the censoring survival estimator in the denominator. This form can be interpreted as a naive estimator counting all in the risk set at time t and each individual in the risk set is weighted by the probability of remaining uncensored until time t.

When two types of events are of interest, a pair of survival times are studied by using their joint survival distribution. Several estimators of a bivariate survival function are proposed in the literature, some are a direct extension of the KM estimator (Dabrowska, 1988). The estimators also depend on the censoring scheme in the bivariate setting; univariate independent censoring where both event times are censored by a common censoring variable, bivariate independent censoring where each event time is censored by its own independent censoring variable and the two censoring variables can be modelled jointly (Akritas and Van Keilegom, 2003; Huang and Louis, 1998; Janssen and Veraverbeke, 2023; Lin and Ying, 1993; Satten et al., 2001). The nonparametric estimation of the joint distribution function of the time to an event of interest and a mark variable has been studied by Huang and Louis, 1998. The authors considered the setting where an event of interest is subjected to right censoring and a mark variable is observed only if the event of interest is not censored. However, they ignored the history of the associated processes in their estimator. On the other hand, a series of consistent and efficient estimators of the survival function of the quality-adjusted life

time have been suggested by Zhao and co-authors (Zhao and Tsiatis, 1997, 1999) where the covariate process history have been utilised in IPCW estimators. On similar lines, various estimators of cumulative medical costs, in the presence of censoring, were developed, using the history of the cost of the treatment (Zhao et al., 2007).

Our interest is in the bivariate survival function of τ and another event related to an associated covariate process. In addition, the censoring event carries information on both types of event and this information needs to be incorporated in the estimation of the bivariate survival function. Because of this special structure, we are in the univariate independent censoring situation. We have proposed an IPCW estimator of this bivariate survival function in Article **III**.
Following are the main assumptions for the application of the IPCW estimation method (Robins and Rotnitzky, 1992, Dong et al., 2020, Zhao and Tsiatis, 1997).

1. IPCW method assumes that the event and the censoring times are independent given the covariate history. In our setting (Article **III**), the time to the discontinuation of treatment and the stopping time related to the integrated covariate process are independent of censoring, which is due to death.

2. To use the inverse of the censoring probability as a weight, it is necessary that this probability is non-zero over the study period. To satisfy this assumption, an artificial end of study is assumed such that the last observation is not censored.

3. Exchangeability of subjects: As discussed in Dong et al., 2020 "given the measured common predictors of the outcome of interest and censoring, censored subjects have the same prognosis for the outcome of interest as do subjects who are not censored." Since in our dataset of wet-AMD the censoring is independent of the event and covariate history, the censored subjects and uncensored subjects have the same disease prognosis. The above assumptions are valid in our case because the disease activity has no direct effect on censoring due to death.

4.1.2 Joint survival estimation

Let $\{Z(t), t \geq 0\}$ be a non-negative covariate process associated with the event of interest τ and let the integrated covariate process be denoted by $\Psi(t)$,

$$\Psi(t) = \int_0^t Z(s)ds.$$

Let $\rho(y)$ denote the time when the trajectory of $Z(.)$ crosses y for the first time and be defined as

$$\rho(y) = \inf\{t : \Psi(t) \geq y\}.$$

Note that $\Psi(t)$ is interpreted as the area under the covariate process $Z(t)$ over the time interval $[0, t]$, and $\rho(y)$ is the first time when the integrated covariate value crosses y. $\Psi(t)$ can also be interpreted as the transformed time to the event with respect to the covariate process. The value of $\Psi(t)$ increases as the follow-up time increases, thereby making it highly dependent on the time of follow-up for each individual. To remove this dependency we use $\Psi(t)$ until a fixed time, say L. The event times are then defined as $\tau = \tau \wedge L$ and $\Psi(\tau) = \int_0^{\tau \wedge L} Z(s)ds$. Such an artificial fixed end of the study is suggested in the literature (Gelber et al., 1989) for reducing the bias in the Kaplan-Meier estimation of transformed time. Gelber et al.(1989) showed that even for independent censoring like administrative right censoring, the transformed time, e.g. the quality-adjusted life time, may induce dependency with the censoring which further induces bias in the estimation of the survival function. To reduce this induced bias, Gelber et al.(1989) suggested introducing an artificial end point to the study before the actual administrative end. Zhao and co-authors (1997, 1999) used such an artificial end of study to satisfy the assumption of the uncensored largest observation for the IPCW methods. Since we have used the IPCW method, the fixed end of the study L was required to satisfy the assumption of IPCW. It is to be noted that in the literature on a quality adjusted life, the transformed time is generally a time with or without toxicity or time before relapse of certain disease. In wet-AMD, if the time with IRF is considered as a time with toxicity then, IRF being a binary variable, this time is actually the area under the IRF trajectory. This reasoning helps to understand the area under the covariate trajectory as the transformed time for all types of covariates.

In the presence of censoring, the integrated covariate process gets truncated due to censoring, hence the event of interest and the event related to the integrated covariate process, both are censored by the same censoring which is independent of both, the event of interest and the covariate process. So, only one censoring indicator is sufficient for both event times. It is to be noted that $\rho(y)$ is finite if $\Psi(t)$ crosses y in a finite time. The observation interval has an upper

bound L, so we define $\rho(y) = \inf\{t \leq L : \Psi(t) \geq y\}$ where $\inf \emptyset = \infty$. Note that $\rho(y)$ is censored by T and we define $\delta_y = \mathbf{1}(\rho(y) \leq T)$ as an indicator function assuming value 1 when $\rho(y)$ is observed and 0 otherwise. In case when $\delta_y = 0$, we use the convention that $\rho(y) = \infty$. This is just to indicate that the integrated covariate process has not crossed the level y in a finite interval $[0, T]$. Further, we assume that the censoring is independent of $(\tau, Z(t), t \geq 0)$.

Since all the formulation is conditional on the covariate process for each i, we treat each process as a non-random function and do not model it. In article **III**, we developed the survival estimator which is consistent for (t, y) such that $P(\tau \wedge C > t \vee \rho(y)) \neq 0$ as follows:

$$S(t,\ y) = P(\tau > t, \Psi(\tau) > y) = P(\tau > t, \tau > \rho(y))$$
$$\simeq \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\mathbf{1}(T_i > t \vee \rho_i(y))}{G(t \vee \rho_i(y))} \right], \tag{4.3}$$

where G denotes the censoring survival distribution. For estimating the censoring survival function in the presence of more than one type of event in the data, various options can be used. We have discussed this question of estimating the censoring survival function in detail in Article **III**. Using Figure 4.1 and Table 4.1, we illustrate the contribution of each individual shown by its trajectory to the numerator or the denominator of the estimator (4.3). The longitudinal trajectories shown in the Figure 4.1 are drawn just for illustration purposes and actual trajectories may not be straight lines. The original time is shown on the X-axis and the integrated covariate process is on the Y-axis.

The joint survival estimators $S(t,\ y)$ in (4.3) can be obtained based on how the sample space shown in Figure 4.1 is divided into four parts. It can be understood on similar lines of the estimator given in equation (4.2) with the risk set in the numerator and the censoring survival function in the denominator.

1. For $y = 0$, $S(t,\ y) = S(t,\ 0)$, we use the information until t as the observed information, that is, the risk set is $\sum 1(T \geq t)$ and the censoring survival function is estimated by using censoring times before t as events. Then the IPCW estimator of $S(t,\ 0)$ is the Kaplan-Meier estimator of survival $P(\tau > t)$ as in equation (4.2). We have shown this equivalence in Article **III**. In this case, the two quadrants on the right contribute to the numerator and also constitute a risk set at time t. The censoring times from the two quadrants on the left contribute as event times in the estimation of the censoring survival distribution in the denominator.

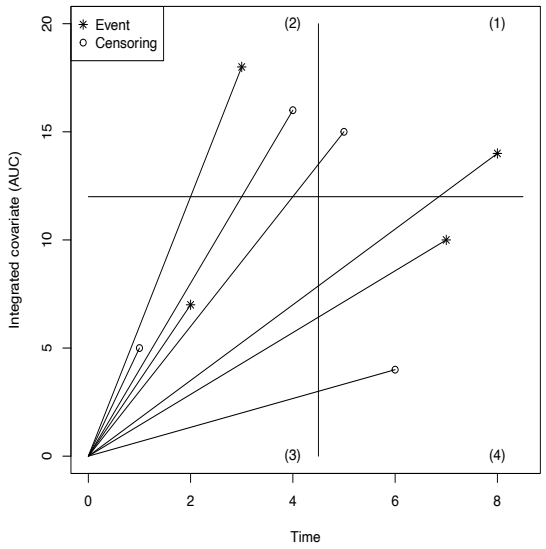


Figure 4.1: The trajectories for eight hypothetical individuals are plotted in this graph of transformed time (AUC) vs. original time. Time is on the X-axis and the integrated covariate value (the area under the covariate trajectory) is on the Y-axis. The solid vertical line near $x = 5$ indicates a fixed value of time t and the horizontal line near $y = 12$ indicates the fixed value of the integrated covariate y. The two dimensional plot area is divided into four quadrants. One event time and one censoring time are plotted in each quadrant for illustration purposes.

2. For $t = 0$, $S(t,\ y) = S(0,\ y)$, we use information until y, that is, the risk set is $\sum 1(T \geq \rho(y))$ or $\sum 1(\Psi(\tau) > y)$ and the censoring survival distribution is estimated by using censoring times before y as events. So, we get an estimator of the survival probability $P(\tau > \rho(y))$ or $P(\Psi(\tau) > y)$. In this case, the observation times in the upper two quadrants contribute in the numerator and also constitute a risk set at y. The censoring from the lower two quadrants contributes as event times in the computation of the censoring survival distribution in the denominator. All four individuals above y contribute in the numerator and none of them contribute as an event time in the estimation of censoring distribution. This estimator is equivalent to the quality adjusted life time estimator given by Zhao and Tsiatis (1997). We have shown this equivalence in Article **III**.

3. For $t > 0, y > 0$, we are interested in the estimation of $S(t,\ y)$, then we use information until t or $\rho(y)$, whichever occurs later, so the risk set is $\sum 1(T \geq t, T \geq \rho(y))$, which is the top right quadrant of Figure 4.1.

The observed information is until $(t \vee \rho(y))$, which is the information from quadrants $(2), (3)$, and (4). The two individuals in the top right quadrant of Figure 4.1 contribute to the numerator. For the denominator, if the censoring survival distribution is not known, it needs to be estimated using the data under analysis. With the data under analysis, there are multiple possible ways to obtain the censoring distribution. It is important to know how each censored observation contributes in each of these possibilities.

(i) **Using external data.** If censoring distribution of the underlying bigger population, from which the analysis data is sampled, is known, then it can be used directly, avoiding the estimation of censoring distribution altogether. This is particularly of interest when the censoring is due to death and can be taken as an independent censoring with regard to the event of interest. In this case, the population mortality distribution can be used. Since the censoring distribution is known and need not be estimated from the observed data, the contribution of the observed data is not applicable as shown in the column (i) of Table 4.1.

(ii) **Using the information only from (T, δ).** The censoring survival distribution is obtained using all individuals in all four quadrants with the censoring as an event. In this case, all the censored times will contribute as event times in obtaining the censoring survival distribution. So, the censored individual in the top right quadrant that is quadrant (1) will contribute to the numerator (risk set) and also to the denominator. All other censored individuals contribute only to the denominator. The censoring indicators for this method for each individual of the example figure Figure 4.1 are given in column (ii) in Table 4.1.

(iii) **Using information from (T, δ) as well as from $(\Psi(t),\ 0 \leq t \leq T)$ and (t, y).** Here, we consider the censoring observed before t in the first dimension or before y in the second dimension which is the censoring observed in quadrants $(2), (3)$, and (4). For this purpose, we reconstruct the data for each fixed pair (t, y) where the event indicator and the observation times are redefined as follows.

$$\widetilde{\tau}_i(t,y) = \tau_i \wedge (t \vee \rho_i(y)),\ \widetilde{T}_i(t,y) = (\widetilde{\tau}_i(t,y) \wedge C_i),$$
$$\widetilde{\delta}_i(t,y) = \begin{cases} 1 & \text{when } C_i \geq \widetilde{\tau}_i(t,y) \\ 0 & \text{otherwise.} \end{cases} \tag{4.4}$$

The censoring survival distribution is obtained using this reconstructed data. In this case, each individual in quadrant (1) in Figure 4.1 contributes to the numerator and censoring observed in quadrants (2), (3), and (4) contributes as event times in the estimation of censoring survival distribution. The censoring indicators for this method are given in column (iii) in Table 4.1.

(iv) **Using the information from** (T, δ) **as well as from** $(\Psi(t),\ 0 \leq t \leq T)$ **and** y**.** Here, we consider the censoring observed before y, that is, in the lower quadrants only. For this purpose, we reconstruct the data for each fixed value of y, where the event indicator and the observation times are redefined as follows.

$$\widetilde{\tau}_i(y) = \tau_i \wedge \rho_i(y),\ \widetilde{T}_i(y) = (\widetilde{\tau}_i(y) \wedge C_i),$$
$$\widetilde{\delta}_i(y) = \begin{cases} 1 & \text{when } C_i \geq \widetilde{\tau}_i(y) \\ 0 & \text{otherwise.} \end{cases} \tag{4.5}$$

The censoring survival distribution is obtained using this reconstructed data. The reconstruction suggested here is similar to the one suggested by Zhao and Tsiatis (1997). In this case, the censored observations in quadrant (2) of the Figure 4.1 do not contribute to the numerator and are not considered as event times for the estimation of the censoring survival distribution in the denominator. The censored observations in quadrant (1) contributes as event to the numerator. All censored individuals in quadrants (3) and (4) contribute as event times in the denominator. The censoring indicators for this method are given in column (iv) in Table 4.1.

The joint survival distributions discussed in this section are useful in comparing multiple groups or multiple treatments or treatment regimes. Two joint survival distributions from two groups can be compared either graphically by plotting their corresponding survival curves or by comparing the percentiles of their corresponding survival distribution. Additionally, one can compare the joint distribution or the marginal survival distribution of the transformed time (which relates to an integrated covariate process) for a specific real time, as well as the marginal survival distribution of the real time, between the two groups. We have implemented this method in Article **III** and demonstrated its use using two examples of wet-AMD data. Among the existing nonparametric bivariate survival estimators, the estimator by Lin and Ying (1993) is closest in the formulation with two related event times and univariate independent censoring however, the

Table 4.1: The contribution of observations and determination of the value of the event $\mathbf{1}(\tau > t, \tau > \rho(y))$. The event indicators are $\delta = \mathbf{1}(T = \tau)$ and $\delta_y = \mathbf{1}(\rho(y) \leq T)$ and $T = \tau \wedge C$. The four quadrants are shown in Figure 4.1. Here $(\tau, C, \rho(y))$ are time-to-event, censoring and stopping time, respectively. Columns (i)-(iv) give four different ways of defining the indicator of the censoring event for estimation of the censoring survival function described in option 3 in section 4.1.2.

Quadrant	(δ, δ_y)	$\mathbf{1}(\tau > t)$	$\mathbf{1}(\tau > \rho(y))$	$\mathbf{1}(\tau > t, \tau > \rho(y))$	(i)	(ii)	(iii)	(iv)
(1)	(1, 1)	1	1	1	-	0	0	0
	(0, 1)				-	1	0	0
(2)	(1, 1)	0	1	0	-	0	0	0
	(0, 1)				-	1	1	0
(3)	(1, 0)	0	0	0	-	0	0	0
	(0, 0)				-	1	1	1
(4)	(1, 0)	1	0	0	-	0	0	0
	(0, 0)				-	1	1	1

censoring may censor only one of the two events. In our setting, although there is univariate independent censoring C for both the events, in reality, τ is censored by C and $\rho(y)$ is censored by $T = \tau \wedge C$. Hence if $\rho(y) < T$, it is possible that C can still censor τ making only one event censored by the univariate censoring; however, if $\tau < C$ and $\rho(y)$ is not observed before τ then C cannot censor $\rho(y)$. In Akritas et al.(2003), a bivariate distribution function of a pair of random variables (X, Y), subjected to independent censoring by (C_1, C_2) was estimated by using the relationship between the bivariate distribution function and the product of the conditional distribution of X given Y and the marginal distribution of Y. The approach to the estimation of the conditional distribution of X given Y was based on the assumption that X is independent of C_2 so that the identity $P(X \leq X \mid Y = y, Y \leq C_2) = P(X \leq x \mid Y = y)$ holds. In our case, this assumption does not hold because $\rho(y)$ can be censored by $\tau \wedge C$. We cannot make use of their estimator of the conditional distribution function or the corresponding smoothed cumulative hazard. These subtle differences made it necessary for the development of the new estimator.

We have developed the estimation methods involving the area under the covariate trajectory. Caution is required while using the area under the trajectory methods. One requirement for using the area under the trajectory as a summary measure is that all individuals in the study should have longitudinal observation at almost the same time and the covariate does not vary much between consecutive visits. We discuss these points in more detail in sections 6.2 and 6.3.

We demonstrated the use of our proposed estimators by applying them to wet-AMD real-world data and wet-AMD clinical trial data in Article **III**. We used

visual acuity (VA) longitudinal trajectories as a covariate process, since VA is measured over a treatment period and is considered as a treatment response. The area under the VA trajectory is commonly used to assess the treatment response as well as to make decision about the treatment of AMD (Gross et al., 2015; Wilding et al., 2012). The event of interest is discontinuation of the treatment due to insufficient response to the treatment. The patients suffering from AMD are elderly, so many dropouts are due to death. However, AMD cannot be a primary cause of death so we considered death as an independent censoring. We applied our proposed estimation method of joint survival probabilities to discontinuation of treatment (as the event of interest) and the area under the VA trajectory (as associated integrated covariate process) in wet-AMD real-world data and to two subgroups of wet-AMD clinical trial data. We showed the use of our method to compare two groups with respect to the event of interest and the associated covariate process. Further, we extended our method for stratified analysis. The severity of the eye disease or health of the eye is generally measured in terms of VA. Higher VA signifies better vision, for example, VA more than 70 is required for a driving license. The integrated VA may be highly dependent on the baseline value of VA. Those who start with, for example, a higher VA attain a higher area under the VA trajectory faster than the others. For this reason, we discussed a stratified estimation for different subgroups created based on baseline VA values. The comparison of the pooled estimate against the unstratified estimate helped to understand the differences in the stratum-specific survival functions. Various lesions in the eye are another important disease marker; for example, AMD treatment targets removal of intra-retinal fluid (IRF). To understand the effect of baseline fluid status on the survival probability, we performed stratified analysis for two strata created based on the baseline binary status of IRF. Our proposed estimators can be easily extended to multiple covariate processes. With consultation with clinicians, one can model such multivariate responses for any chronic diseases.

In chronic diseases, patients require treatment for a considerably longer period. During a long follow-up, the patient's situation may change in terms of available physical help, financial situation, and the treatment response pattern. All these factors lead to missing clinical visits and may result in under-treatment or termination of treatment either by the patient or by the clinician. Recognising the treatment response pattern until that time is important in order to help make the decision about continuing the treatment. The discontinuations due to the reasons not related to the disease or to the treatment can be incorporated as independent censoring. The decision to continue the treatment at each treatment visit is essential in the case of chronic diseases. I discuss this more in the next

section.

4.2 Medical decision making

Medical professionals treating chronic diseases need to make a decision at each patient visit about the future treatment, taking into account the patient's characteristics, the history of the treatment given to the patient and the patient's response to the treatment. They can utilise the information available to them from other patients that share the characteristics or the history of the patient in question. The decision to continue the ongoing treatment or to switch to a different treatment has consequences not only for the patient but may also affect the decision making of the whole healthcare system at the local or the national level (Masic, 2022). Some authors distinguish between population based decision making and personalised decision making (Mueller and Pearl, 2023), while others consider the decision of choosing the optimal treatment regime for each patient as a set of rules (Hager et al., 2018). Cook and Lawless (2018) discuss individual level decision making by analysing the process related cost and benefits in multistate model settings.

We developed a nonparametric decision making methodology based on the covariate history of the patient. We formulated the problem of discontinuation of the treatment, discussed in the previous chapter, in the conditional setting. We were interested in the estimation of the hazard of discontinuation, given the integrated covariate process for each individual. For this purpose, we proposed the estimation of the conditional cumulative hazard function of the discontinuation of treatment, conditional on the events related to the integrated covariate process in Article **III**. The estimation of the hazard rate in nonparametric survival regression models has been discussed in the literature (Van Keilegom and Veraverbeke, 2001; Van Keilegom et al., 2001). The approach commonly seen is of first estimating the hazard rate and then smoothing it to get the cumulative hazard or survival function estimator. We wanted to develop a simple mechanism that can be used by clinicians for decision making in their regular clinical visits and both our datasets had regularly measured covariate values, so we did not follow the route available in the literature and decided to estimate the cumulative hazard directly,

4.2.1 Role of conditional hazards and estimation

The decision to discontinue treatment depends on the covariate history. By applying the formulation from the previous section, we can derive the conditional hazard function. This function describes the likelihood of observing an event

based on the covariate history or the integrated covariate process. We have discussed such conditional hazard estimation in detail in Article **III**.

The Nelson-Aalen type estimator of the conditional hazard of τ given $\Psi(t) \leq y$ using the data $(T_i, \delta_i, \rho_i(y), t \leq T_i)$, $i = 1, \ldots, n$ is given as follows.

$$\widehat{\Lambda}(t \mid t \leq \rho(y)) = \sum_{T_i \leq t} \frac{\mathbf{1}(T_i \leq \rho_i(y))\delta_i}{\mathbb{R}_1(T_i)}, \tag{4.6}$$

where the risk set $\mathbb{R}_1(t) = \sum_j \mathbf{1}(T_j \geq t)\mathbf{1}(t \leq \rho_j(y))$, consists of those who have not yet experienced the event and the integrated covariate trajectory has not crossed y by time t. Contribution of each individual in the estimation of a conditional hazard can be shown clearly using Figure 4.1 and Table 4.1 on lines similar to the joint survival function. We have discussed this in detail in Article **III**.

Conditioning on one or more covariate processes, the estimate of the hazard of observing an event of interest can be obtained as shown in Article **III**. The decision making process can be made transparent by studying various conditional hazard graphs as shown in Figure 4.2. In this figure, we present the estimates of the cumulative hazard functions conditional on two integrated covariate processes - integrated visual acuity (VA) over time and the total number of injections. For example, at 40 weeks, the hazard estimate of discontinuation of the treatment, conditional on the number of injections > 6 and the integrated VA ≤ 3000 is lower than that for the number of injections ≤ 6 and the integrated VA ≤ 3000. But, at 80 weeks, the hazard estimate of discontinuation of the treatment, conditional on the number of injections > 6 and the integrated VA ≤ 3000 is much higher than that for the number of injections ≤ 6 and the integrated VA ≤ 3000. This indicates that with more injections, the response (in terms of the integrated VA) remains low, then the chances of discontinuation of the treatment is higher than the lower number of injections and the same response. All four graphs show this trend of the hazard increasing initially and then stabilising for a lower integrated VA with a less number of injections and the trend of the hazard being lower initially and then increasing for a lower integrated VA with a high number of injections. The two graphs in the first row with the condition on six injections show that the hazard of more than six injections becomes higher than the hazard of less than six injections in the later part around 60 weeks. In the second row, for 12 injections, the trends of these two types of hazard are similar, as the hazard of more than 12 injections increases with time, while the hazard of less than 12 injections increases initially but stabilises later.

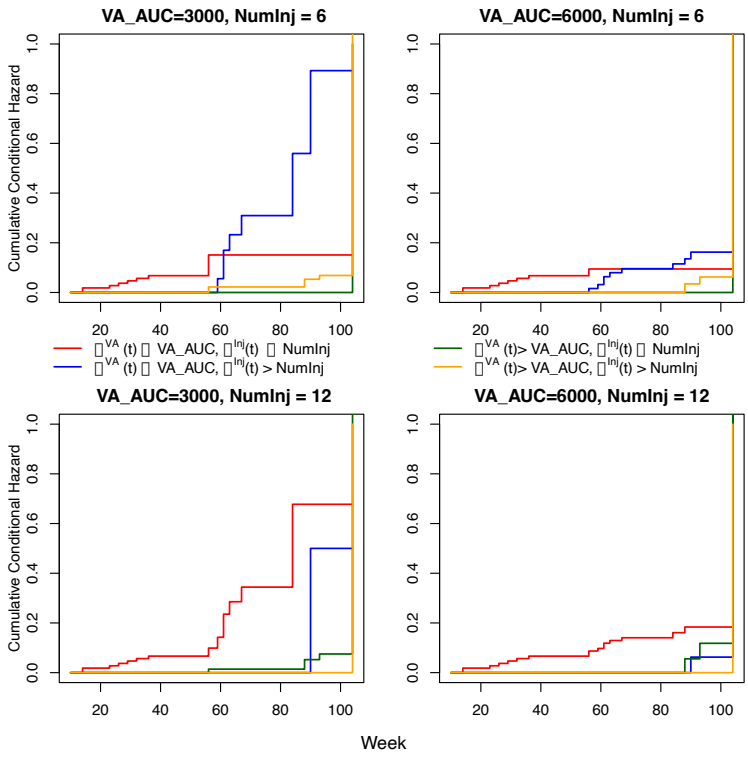


Figure 4.2: Conditional cumulative hazard plot. Time is plotted on the X-axis. Each subplot shows conditional cumulative hazard functions plotted for four combinations of conditions on integrated VA value and the total number of injections until time t. For each subplot, the title shows the values of the integrated VA and the total number of injections used for conditioning combinations.

4.3 Comments on proposed estimators

In Article **III**, we proposed an estimator of the joint survival function of the time to event of interest and the integrated covariate process by using all the available information on the covariate process over the observation period. Our proposed estimators for fixed (t, y) have been developed for the integrated covariate process; however, they can also be used for the estimation of the survival function of, for example, the quality-adjusted survival time or the total medical cost. If the censoring distribution is known from the underlying population and if the assumption of independent censoring is valid, the proposed estimator with known censoring distribution gives an unbiased estimation of the bivariate survival function $S(t, y)$. In settings where the censoring is due to reasons unrelated to the event of interest and the covariate process, for example, death, in our dataset,

we can use registry data from the general population from which the patient population comes, in order to obtain the censoring distribution.

In terms of computational complexity, the computation time required for the estimation of the survival function using reconstructed data is more than 5 times, as compared to the KM censoring from the observed data, even for the small sample size of 25 and the variance estimation takes even longer. In our case, the censoring mechanism is independent of τ and the covariate process, and so, the data reconstruction method has a little advantage over other methods. When the censoring mechanism depends on the covariate process, we think that the data reconstruction method will bring in the visible effect of y through the estimate of the censoring distribution, making it preferable over other methods.

Our proposed estimators of the joint survival function and the conditional survival functions can easily be extended to multivariate settings. When two covariate processes are of interest, we define two integrated processes $\Psi^{(1)}(t)$ and $\Psi^{(2)}(t)$, and extend the estimation methodology. Conditional survival probabilities are especially helpful in decision making regarding the change in treatment regime. For a new patient or for an ongoing study, the decision to continue (or discontinue) the ongoing treatment regime at any time during the follow-up can be made using the existing joint or conditional distribution from the existing studies where the patient population is the same as in the new study, and the covariate measurement settings are also similar in the two studies.

Chapter 5

Article summaries

After detailed elaboration of statistical methods in previous chapters, I present short summaries of the three research articles in this chapter.

5.1 Article I

The objective of the first article was to study various alternative methods that give finite parameter estimates in the case of quasi-separation in the count data with respect to parameter estimation and outcome prediction. In this first article of this thesis, the event counts were modelled by using Poisson distribution. The Poisson regression model was reviewed with special focus on non-existence of Maximum likelihood (ML) estimates in a quasi-separation case shown in Figure 2.1. Firth's correction to obtain a finite parameter estimate was explored in detail. Its prediction properties for the Poisson regression model were also examined. Firth's method is known to shrink the parameter estimates towards null, in order to get the bias reduced estimates. This bias reduction in the estimation of the regression coefficients introduces bias in the intercept estimate, adversely affecting the prediction of the response. To avoid this undesirable effect on the prediction, two modifications to Firth's method, 'FL with intercept correction' (FLIC) and 'FL with added covariate' (FLAC), that gives unbiased predictions, proposed by (Puhr et al., 2017), were studied in this article. Although these two approaches, FLIC and FLAC, give different results for the logistic regression, it was observed from the results in this article that all three methods, FL, FLIC and FLAC, the estimates for all the covariates except the intercept term coincide for Poisson regression. The proof of these equal parameter estimates was given in the appendix of the article. It was proved that the intercept term in FLIC and FLAC estimation also coincides.

The results from the exact conditional method were used for parameter es-

timation but not for prediction because none of the existing software provides an estimate for the intercept term. The Bayesian augmented data method with weakly informative priors was also presented in this article. The Bayesian data augmentation method gave comparable unbiased estimates to those of FLIC (and FLAC). However, the Bayesian data augmentation method required appropriate prior information, which was not a requirement for FLIC and FLAC.

The advantages of Firth's penalised likelihood method were discussed in detail in this article. The non-existence of ML estimates, especially due to separation, was the central problem of this research. Two real-world examples with rare events data were discussed. A simulation study based on one of these datasets was performed for continuous, binary and categorical types of covariates.

Appropriate methods of interval estimations were used and demonstrated with figures. In this thesis summary, these methods are already listed in section 6.1

5.2 Article II

The objective of the second article was to obtain a joint inference from multiple disease markers in the form of simultaneous credible regions. In this article, we developed simultaneous credible regions from Bayesian posterior distributions of multiple binary markers. Getting the joint Bayesian posterior distribution for multiple markers was the key point of this article. If a disease marker is used in any of the previous studies, it is possible to compile all previous information into a prior distribution of the natural parameter of the marker, independent of the effect measure used in previous studies. It was demonstrated that once the joint posterior for all markers was available, one can easily get the posterior distribution of the required effect measure. An equal tailed interval (ETI) and highest posterior density intervals (HDI) were constructed for up to four markers. An algorithm to compute the intervals for multiple markers was described in the article and the corresponding R-shiny application was made available publicly. Please note that there is a typo in Step 1 of the algorithm of the ETI computation in section 3.1, the starting value of α_{min} has to be $\alpha/4$ or a smaller positive number.

For a single binary marker, the proposed methods of interval computation, ETI and HDI, were compared with the existing methods such as Wald, Agresti-Caffo, Score (Miettinen-Nurminen), Newcombe, and exact nonparametric. The methods of interval estimation at higher dimensions were scarcely available, so, a thorough simulation study was performed to study the properties of ETI and HDI for multiple markers. Various priors like non-informative prior, weakly in-

formative priors and highly informative priors were used to obtain simultaneous posterior intervals. The effect of these priors was studied for equal and unequal sample sizes in two groups. With an example dataset, it was shown that for multiple markers, the method of simultaneous credible regions can be utilised for gatekeeping procedures, similar to those in clinical trials. The article also discussed the differences in the inference drawn from the joint posterior distribution of all markers against the marginal posterior of each individual marker. The case study of the wet-AMD data was discussed at length. Posterior simultaneous credible regions for multiple markers were used to compare two subgroups. Another case-study of a clinical trial for a treatment on migraine is discussed. The emphasis of the article was on obtaining a common inference from multiple markers, so, some other aspects, like the length of the interval in the case of a single marker, or the area (size) of simultaneous regions for multiple markers were not discussed in this article.

5.3 Article III

The objective of the third article was to study jointly an event of interest and associated integrated covariate processes. This article discussed the nonparametric joint estimation of survival of two types of events. During the treatment of chronic diseases, the treatment response over the follow-up period can be considered as a covariate process. Such a covariate process, which is a treatment response, is associated with the event of interest. Our main aim in this article was to estimate the joint survival probability of the time to event of interest and the stopping time of the integrated covariate process. In real-world settings, the time to event is often censored. Under the assumption that the censoring is independent of the time to event and the covariate process, we proposed the inverse probability of the censoring weighted estimators of the joint survival function.
If the censoring survival distribution from the underlying population is known then the unbiased joint survival estimate of the two events mentioned above can be obtained. If the censoring survival distribution is not known, the joint survival estimators and their efficiency differ according to the handling of the censoring distribution. Two treatments or two treatment regimes can be compared either graphically by plotting their corresponding survival curves or by comparing the percentiles of their corresponding survival curves.
We illustrated our methods using two sets of data on age-related macular degeneration of the eye, one from a clinical trial, and the other from a real-world study. In case of the wet-AMD, the event of interest was the time to the discontinuation

of the treatment and the other event was the stopping time of the integrated covariate process (the first time when the trajectory crosses a fixed point). In order to examine the effect of the baseline covariate values, a stratified analysis was performed.
Further, in order to choose the best suited treatment at each treatment visit, the conditional cumulative hazard function of the time to event, given the history of the covariate process, was developed. This method can help clinicians treating chronic diseases in the decision making of treatment change or discontinuation in their regular clinical visits.

Chapter 6

Discussion

In this thesis, I discussed statistical methods and their synthesis in the context of the event data observed in the medical field. The proposed methods were applied to both, randomised clinical trials and observational studies. The first article examined parameter estimation methods for count data in regression analysis. The frequentist parametric and semiparametric methods, along with some Bayesian parametric methods were compared with an emphasis on the likelihood penalisation method. A likelihood penalisation method for GLMs suggested by Firth (Firth, 1993) is very well studied for binary logistic regression. Predicted responses obtained by using the parameter estimates of the Firth penalised likelihood method are biased upwards due to the bias in the intercept estimate. Improvements such as FLIC, FLAC in the Firth method for reducing the bias in the prediction are also frequently used for binary logistic regression. In Article **I**, these improvements in the Firth penalisation method were extended to Poisson regression. The most important finding was that the covariates other than the intercept have the same estimates using the Firth method, the FLIC method, and the FLAC method. Also, the intercept estimates for FLIC and FLAC coincide. This was not seen for binary logistic regression for which these methods had initially been developed.

The motivation for Articles **II** and **III** was an observational study on the wet-AMD, involving multiple markers and their joint impact on decision making. In Article **II**, the joint posterior distribution of multiple binary markers was studied in the framework of contingency tables with the main purpose of comparing two groups of patients. The biggest advantage of the proposed method is in the combined inference obtained from various effect measures. The method was useful in comparing two groups of patients based on the joint posterior distribution of the natural parameters related to multiple markers, or in comparing the joint poste-

rior distribution at two time points. Article **II** developed two types of credible regions for effect measures of multiple markers- the equal tailed region and the highest density region.

These highest density regions may not be equal tailed and may not be connected. The main contribution of this article was the development of an algorithm to estimate the simultaneous credible regions for multiple markers. Another important contribution of this research was the development of the gatekeeping procedure in terms of simultaneous credible regions. Along with this article, a user friendly R-shiny app was also provided to get the joint cumulative distribution of multiple markers.

In Article **III** the data from two follow-up studies, the CATT clinical trial and an observational study, were considered. It discussed the effect of longitudinal markers on the event of interest. Using the method of inverse probability of censoring weights, the joint estimation of the survival function of an event of interest and an event related to an integrated covariate process was proposed in a nonparametric way. The effect of the method of estimating the censoring distribution on the efficiency of the estimator was discussed in detail. In this article, the cumulative hazard of observing an event of interest conditional on the covariate process history was the instrument used for medical decision making.

6.1 Variance computation and interval estimation

All three articles in this thesis discuss variance computation or interval estimation at length. In the first article, although the focus is on the point estimation of the regression parameter, sufficient attention is given to interval estimation. For each method of point estimation, the appropriate interval estimation is discussed. For maximum likelihood estimation, a Wald type of confidence interval is computed. For the exact permutational test, the exact confidence intervals obtained from the permutational distribution are used. The exact intervals are generally conservative, so, the mid-p correction is also provided. In mid-p correction, half the exact point probability is subtracted from the tail probability to get the modified p-value. The mid-p adjustment reduces the discreteness of the parameter distribution and thereby decreases conservatism and reduces the interval length. Although the $100(1-\alpha)\%$ level is not guaranteed theoretically, many simulation studies have shown that the mid-p correction does not compromise the coverage level of $100(1-\alpha)\%$ (Hirji et al., 1991).
The Firth penalisation method is a shrinkage method. So, in addition to reduc-

ing the bias, it also reduces the variance in the parameter estimation. For this method, the profile likelihood based confidence intervals are reported, which are recommended when the likelihood is not necessarily symmetric about the MLE. Figure 6.1 shows the difference in the Wald and profile likelihood based interval. For the Bayesian data augmentation method, the effect of prior information on the parameter estimates was also studied by varying the prior.

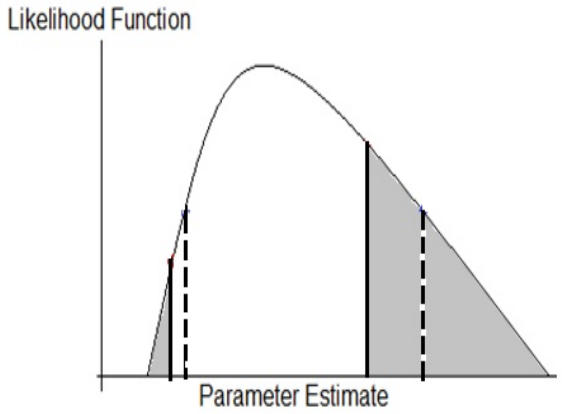


Figure 6.1: The Wald and profile likelihood based interval for a non-symmetric likelihood function. The interval limits for the Wald interval are shown by black solid lines and the region beyond these limits is shown in grey colour. The interval limits for the profile likelihood based interval are shown by dashed lines.

The second article is about Bayesian posterior credible intervals. For multiple markers, the posterior region was obtained from the joint posterior distribution. Two types of credible intervals, equal tailed interval (ETI) and highest density interval (HDI) were proposed in Article **II**. These Bayesian posterior intervals ETI and HDI have a similar interpretation respectively, to that of the wald type of interval and to the profile likelihood based interval among frequentist confidence intervals. ETI can be viewed as a method similar to that of wald type of interval where the interval limits are equidistant from MLE of the parameter distribution while HDI can be viewed closer to profile likelihood based interval in which the interval consists of points with their density above a certain threshold rather than their distance from the MLE. A detailed simulation study was conducted to study the effect of the prior on the posterior credible regions. For single and multiple markers, the credible intervals were reported for many combinations of the prior information and the effect measures of the markers. The algorithm to compute these credible regions was given in the article and the corresponding R-shiny code was made available. The use of credible regions for the gatekeeping procedure

was illustrated with an example.
In Article **III**, pointwise nonparametric joint survival estimation is discussed. The article also discussed the asymptotic theory and variance estimation. However, asymptotic variance-covariances are not always available in the closed form and some estimators of the quantities involved are needed. Hence, resampling methods such as jackknife or bootstrap are recommended. Since the methods developed in this article are nonparametric, the variance computation using nonparametric resampling methods like the jackknife or bootstrap is justifiable. The advances in computer hardware and software have made running these computationally heavy resampling methods much easier in recent years than before.

6.2 Computation of the area under the trajectory

We used wet-AMD real-world data and wet-AMD clinical trial data to illustrate the methods based on the area under the covariate process or an integrated covariate process. To make use of the integrated covariate process, it is necessary that all individuals have measurements with the same frequency and the covariate value changes slowly and continuously between two measurements (Bonate, 1998). Methods involving an area under the curve become biased if there are missing data or imbalanced data (Bell et al., 2014; Wilding et al., 2012). The wet-AMD clinical trial data have measurements at every four weeks and the wet-AMD real-world data have measurements approximately every month, not exactly on the same day of each month but a day near enough. The visual acuity of the eye is a slowly changing response and it is not expected to change suddenly within a few days. Hence, the observations made approximately every month are acceptable. We used the actual day of the measurements and converted all time measurements into weeks from the day of the entry of the person into the study. Since both our datasets had almost regular measurements, we connected consecutive measurements by lines and then computed the area under such piecewise linear curve using the trapezoid rule of integration. Other more involved smoothing techniques, like spline interpolation, kernel smoothing, or the broken stick model can be employed, but were not necessary for the illustration of our methods, so, they were avoided. However, if such regular measurements are not available, one has to model the trajectory on some other independent covariates and obtain imputed values to get regularly spaced observations for all individuals in the study. If the interest is in the expected cumulative response

for a group of patients over the treatment period, then various approaches used in estimation of the population mean in the presence of right censoring can be employed (Datta, 2005). Even though the methods discussed in this section were developed for wet-AMD in mind, they can be modified for other chronic diseases by incorporating the domain knowledge. The discussion of a significantly large number of missing visits at random or any missing visits not at random are not in the scope of this thesis.

6.3 Real-world clinical data and biases

In the introduction to this thesis, it was mentioned that the analysis methods used in the randomised clinical trials may not be applicable to observational studies. There are many differences in these two types of studies, which need to be taken into account while making a choice of design, data collection, analysis methods, and interpretation. I briefly discuss these points in this section. In our case study of wet-AMD, many of these issues did not occur, so they were not discussed earlier. Since clinical trials and observational studies have different objectives, it should be reflected in their respective designs. For example, the review article by Colditz (2010) discusses the strengths and limitations of observational study designs compared to randomised clinical trials, especially involving dietary data. The article discusses that dietary exposure is population specific, making the dosing of new treatment in regular clinical visits to be tailored for that population. The next important aspect is data collection and deciding on important endpoints or markers. Observational studies can have much longer follow-ups than clinical trials, making it possible to study many repeated observations of markers, and also some rarely observed markers. Hence, the selection of appropriate primary endpoints for observational study is considered important (Gómez et al., 2016). Ording et al. (2016), in their article on the translation of endpoints from clinical trial to observational studies, discuss the importance of choosing the appropriate multiple endpoints in clinical practice, which may not be studied in a clinical trial with limited follow-up and stringent inclusion criteria. In observational studies, the confounding effect by some other variables on both the covariates of interest and the response, need to be properly understood (Meuli and Dick, 2018) to avoid misinterpretation of the confounding effect as a positive treatment effect or harm. The dose-response relationship in a real-world setting is often far from the one observed in clinical trials for various reasons. For example, in connection with the treatment of wet-AMD, many researchers (Nair et al., 2022) expressed their concern about fewer injections in clinical practice, leading to a higher rate of

irreversible vision loss. They gave a detailed description of the current status and the problems related to wet-AMD treatment. There is clearly a need for further research in new drug development as well as optimal treatment intervals in wet-AMD treatment. In such cases, the estimation target is the longest duration between the injections to maintain the desired eye condition. In clinical practice, the same clinic may not treat a patient from the onset of the disease until the event of interest. This can create late entry bias or incomplete follow-up (Bull and Spiegelhalter, 1997). Also, the baseline measurement is not necessarily disease free. To overcome this issue, many studies perform some baseline adjustment while analysing the data. However, such baseline adjustments are not always beneficial or correct (Glymour et al., 2005).

The use of real-world data to assess the effectiveness of the treatment may cause the so-called immortal time bias. In our case study of wet-AMD this issue does not arise because the follow-up of each patient starts when the first treatment dose (injection) is administered. This being an invasive procedure, it coincides with a clinical visit. Hence, the information is available in the health register. We have observed that the start date of the treatment is mostly within two weeks of the date of diagnosis of wet-AMD. Of course, a patient begins to respond to a treatment only after it is first administered (and not when the disease is diagnosed). The patient population is self-selected in the real-world setting. Our purpose was to understand the dependency between different markers and the association with the time-to-event in a specific patient population. Hence, selection bias is not a cause for concern. When comparing treatments, regimes, or healthcare providers, selection bias will become a cause for concern in a real-world setting.

In this thesis, I used the wet-AMD data from a clinical trial and from an observational study. The data came from two different populations and, more importantly, it was not clear whether the measurements of VA and other lesions were performed with the same setting and the same procedure. Due to these differences, I did not combine these two datasets. However, methods are available for combining multiple datasets in certain situations. The biases arising from such combined datasets and the advantages of combining the datasets are discussed by Bareinboim and Pearl (2016) in a causal inference setting.

6.4 Possible extensions

In this section, I sketch out the possibilities for further research. The first two articles of this thesis discuss the modelling of specific types of discrete outcomes,

such as count data or multidimensional binary data. The methods developed for these data types can be extended to other data types. The methods discussed in Article **I** are now used for the binary logistic and Poisson log-linear model. They can be extended to other GLMs and their properties can be studied in more detail. Similarly, the methods developed in Article **II** for multiple binary markers can be extended to other data types or mixed data types. Although these extensions are important and will enrich the field of statistical estimation methods, here I have focused more on the methods that will help in medical decision making.

Informative and non-informative censoring

In Article **III**, we have considered only one type of censoring, which is independent of the event of interest and the covariate processes. However, in many situations, it is possible to have multiple reasons for right censoring, which are either dependent or independent of the disease and the treatment. Here we mention the extension of our method to such cases. As before, let τ be the time to the event of interest, which, in the wet-AMD example, is the discontinuation of the treatment (the decision made by the clinician on account of complete blindness or a stage beyond cure). Let the two types of censoring, one denoted by T_{di} be the informative dropout or dependent censoring (dropping out due to side effects of the treatment) and the other denoted by T_{ci} be the non-informative dropout or independent censoring (death of the patient or the end of the administrative follow-up), and let $C_i = \min(T_{di}, T_{ci})$. Along with the event indicator, we also observe the type of censoring. The data can be redefined as

$$D_i = \{X_i = \min(\tau_i, T_{di}, T_{ci}), \Delta_i, Z_i(t) : 0 \leq t \leq X_i\}, i = 1, \ldots, n$$

where

$$\Delta_i = \begin{cases} 1 & \text{for } X_i = \tau_i \\ 2 & \text{for } X_i = T_{di} \\ 0 & \text{for } X_i = T_{ci} \end{cases} \tag{6.1}$$

In case of both, the informative and the non-informative censoring, one can consider the competing risk approach (Dong et al., 2020). If the two types of censoring are independent of each other, then another option is to replace the weight in the denominator in equation (4.3) by the product of the two censoring survival functions: $G_c(.)$ for independent censoring and $G_d(.)$ for dependent censoring. The distribution of independent censoring may be taken from an external source and need not be estimated from the data.

Area under the curve using the natural progression of disease

To compare two patients or two groups of patients with respect to the integrated covariate process, we need to estimate their integrated covariate processes over a fixed interval of time. Let C_A be the administrative end of the follow-up time. To estimate the integrated covariate process or area-under the curve over the interval $[0, C_A]$, the covariate process needs to be observed until C_A. In the case of wet-AMD, if censoring C due to death is observed, the required area-under the curve is to be computed until C since after death the response trajectory is non-existent (Kurland et al., 2009). If the patient has reached the stage of blindness at certain known time, then also the VA becomes zero after that time point. In principle, the cumulative response up to the administrative follow-up time or death or blindness is needed for comparison purposes. We continue to use the notation C_A to indicate the minimum of the three time points. If the event τ of treatment discontinuation is observed during follow-up $[0, C_A]$, soon after treatment has stopped, in the absence of any treatment, the disease will progress as per its natural progression. So, to obtain the area after the event time, the area under the trajectory of the natural progression of the disease can be used. So, the area under the covariate process until the administrative end of the study can be estimated as

$$\int_0^{C_A} Z_i(s)ds \simeq \Psi_i(\tau_i) + \int_{(\tau_i+\epsilon_i)}^{C_A} g(t; a, b)dt,$$

where $g(t; a, b)$ is a carefully chosen known function, essentially capturing the natural progression of the disease with regard to $Z_i(.)$. For example, one such function suggested by Shah and Del Priore (2007) based on the meta-analysis of many wet-AMD studies is

$$g(t; a, b) = Z_i(\tau_i) - \frac{t - \tau_i}{a(t - \tau_i) + b}, \text{ with } a = 0.0133, b = 0.1448.$$

AUC as the difference in the observed and the natural progression

The repeated measurement data obtained during multiple clinical visits can easily be decades-long longitudinal data. Generally, the trajectories of the most important outcome are considered for modelling. The most common effect found in the literature is the loss or gain in the response compared to the baseline measurement, for example, VA in the case of wet-AMD. Age-related diseases need continuous treatment from the time of the diagnosis of the disease until the end of life of the patient, or until full recovery, which is rare. To correctly estimate the

effect of any treatment, one must also study the natural progression of the disease, had no treatment been given to the patient. In clinical studies, the time of enrolment is not necessarily the time of the onset of the disease. So, the baseline measurement is not the disease-free measurement. Another important consideration is the progression of the disease over time. The effect of the treatment can be considered in the form of the area between the two trajectories; the observed trajectory when treatment was administered, and the other trajectory of possible natural progression of the disease if no treatment had been administered, as shown in Figure 6.2. However, only one of the two trajectories can be observed. This is similar to counterfactuals in causal reasoning (Pearl, 2000).

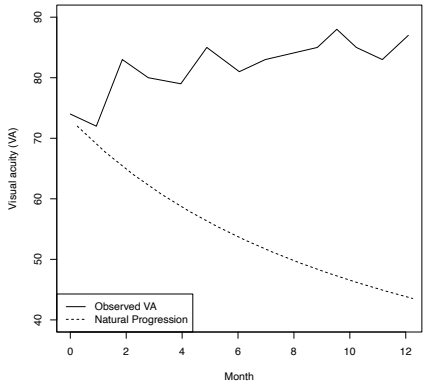


Figure 6.2: The observed VA trajectory for one patient under treatment and possible VA trajectory under natural progression of wet-AMD disease when not treated.

In Figure 6.2, the observed VA trajectory of one patient is drawn. The corresponding natural progression of the disease, starting from the patient's baseline VA value, is plotted only for illustration purposes. The natural progression trajectory is obtained by using the formula given in the meta-analysis article (Shah and Del Priore, 2007). The problem where baseline measurements affect future responses can be attenuated by considering the difference in the observed trajectory and the possible natural progression trajectory. For a better understanding of the disease activity and the treatment effect, it is better to measure the area between the two trajectories, the observed outcome trajectory and the natural progression trajectory, rather than measuring the area under the observed trajectory.

6.5 Concluding remarks

The gist of this research is in the utilisation of the available information to the fullest. The thesis emphasises the practice of critically examining the validity of the assumptions underlying statistical methods and understanding their limitations when analysing research datasets, and suggests appropriate alternative methods when the assumptions of the classical methods are not met. In small sample or rare events study of univariate marker, to avoid the problem of small sample bias in ML estimation, one should use the Firth correction method. For predicting the outcomes, one of the modifications to the Firth method, FLIC and FLAC, should be used. In multiple marker settings, when the treatment effect is understood in terms of multiple markers together, the inference should be based on the combined analysis of all markers. Hence, one should rely on methods which provide a combined inference from all markers such as simultaneous credible regions. Such approach captures the dependence structure among multiple markers. Lastly, in the time-to-event setting with longitudinal data on the associated covariate or marker processes, the estimation of the joint survival probabilities of the event of interest and stopping times related to the covariate processes using the IPCW approach provides a flexible nonparametric approach. This approach avoids making distributional assumptions related to the dynamics and dependence structure of the covariate processes. Hence, it is a very useful approach in medical decision making. Overall, the factors including the intervention, its mechanistic understanding (even partial) and cost, clinical outcomes, and patient-reported outcomes (when available) should be at the forefront of any medical decision making.

Bibliography

Abramovich, F., & Angelini, C. (2006). Bayesian maximum a posteriori multiple testing procedure. *Sankhy¯a : The Indian Journal of Statistics*, *68*(3), 436–460.

Aerts, M., Claeskens, G., & M.P., W. (2002). Some theory for penalized spline generalized additive models. *Journal of Statistical Planning and Inference*, *103*(1), 455–470. https://doi.org/https://doi.org/10.1016/S0378-3758(01)00237-3

Agresti, A. (2013). *Categorical data analysis.* (3rd ed.). Wiley-Interscience.

Akritas, M., & Van Keilegom, I. (2003). Estimation of bivariate and marginal distributions with censored data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *65*(2), 457–471. https://doi.org/10.1111/1467-9868.00396

Albert, A., & Anderson, J. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, *71*(1), 1–10. http://www.jstor.org/stable/2336390

Al-Youssef, M., Taschwer, K., & Thaler, S. (2020). Stichprobentests sollen klarheit bringen. *Der Standard. 2020, April 3.*

Andersen, P., Borgan, Ø., Gill, R., & Keiding, N. (1992). Statistical models based on counting processes. In *Springer series in statistics*. Springer.

Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences - PNAS*, *113*(27), 7345–7352.

Bell, M. L., King, M. T., & Fairclough, D. L. (2014). Bias in area under the curve for longitudinal clinical trials with missing patient reported outcome data. *SAGE open*, *4*.

Berger, J., Bernardo, J., & Sun, D. (2015). Overall objective priors. *Bayesian Analysis (2015),10, Number 1, pp. 189–221.* DOI:%2010.1214/14-BA915

Berk, R., & MacDonald, J. M. (2008). Overdispersion and poisson regression. *Journal of quantitative criminology*, *24*(3), 269–284.

Bonate, P. L. (1998). Coverage and precision of confidence intervals for area under the curve using parametric and non-parametric methods in a toxicokinetic experimental design. *Pharmaceutical research*, *15*(3), 405–410. https://doi.org/https://doi.org/10.1023/a:1011968129921

Bull, K., & Spiegelhalter, D. J. (1997). Tutorial in biostatistics survival analysis in observational studies. *Statistics in medicine*, *16*(9), 1041–1074.

Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data* (1st ed., Vol. v.Series Number 30). Cambridge University Press.

Chakravarthy, U., Evans, J., & Rosenfeld, P. (2010). Clinical review, age related macular degeneration. *British Medical Journal 2010, 340, c981.*

Chi, Y. Y., & Ibrahim, J. G. (2005). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, *62*(2), 432–445. https://doi.org/10.1111/j.1541-0420.2005.00448.x

Colditz, G. A. (2010). Overview of the epidemiology methods and applications: Strengths and limitations of observational study designs [PMID: 21132580]. *Critical Reviews in Food Science and Nutrition*, *50*(sup1), 10–12. https://doi.org/10.1080/10408398.2010.526838

Cook, R. J., & Lawless, J. F. (2018). *Multistate models for the analysis of life history data*. CRC Press.

Correia, S., Guimarães, P., & Zylkin, T. (2019). Verifying the existence of maximum likelihood estimates for generalized linear models. *arXiv: Econometrics*. https://api.semanticscholar.org/CorpusID:85459713

Cox, D. R. (1970). *Analysis of binary data*. Chapman & Hall, London.

Cox, D., & Hinkley, D. (1974). *Theoretical statistics (1st ed.)* Chapman; Hall/CRC. https://doi.org/https://doi.org/10.1201/b14832

Cuparić, M. (2021). Asymptotic properties of inverse probability of censored weighted u-empirical process for right-censored data with applications. *Statistics*, *55*(5), 1035–1057. https://doi.org/10.1080/02331888.2021.1998054

Dabrowska, D. M. (1988). Kaplan-meier estimate on the plane. *The Annals of Statistics*, *16*(4), 1475–1489. https://www.jstor.org/stable/2241775

Datta, S. (2005). Estimating the mean life time using right censored data. *Statistical Methodology*, *2*(1), 65–69. https://doi.org/https://doi.org/10.1016/j.stamet.2004.11.003

Datta, S., Bandyopadhyay, D., & Satten, G. (2010). Inverse probability of censoring weighted u statistics for right censored data with an application to testing hypotheses. *Scandinavian Journal of Statistics*, *37*(4), 680–700. https://doi.org/10.1111/j.1467-9469.2010.00697.x

De Gruttola, V., & Tu, X. M. (1994). Modelling progression of cd4-lymphocyte count and its relationship to survival time. *Biometrics*, *50*(4), 1003–1014. http://www.jstor.org/stable/2533439

Den Teuling, N. G. P., Pauws, S. C., & Van den Heuvel, E. R. (2023). A comparison of methods for clustering longitudinal data with slowly changing trends. *Communications in Statistics - Simulation and Computation*, *52*(3), 621–648. https://doi.org/10.1080/03610918.2020.1861464

Dewan, I., Deshpande, J. V., & Kulathinal, S. B. (2004). On testing dependence between time to failure and cause of failure via conditional probabilities. *Scandinavian Journal of Statistics*, *31*, 79–91.

Dmitrienko, A., & Tamhane, A. C. (2007). Gatekeeping procedures with clinical trial applications. *Pharmaceutical Statistics*, *6*(3), 171–180. https://doi.org/https://doi.org/10.1002/pst.291

Dong, G., Mao, L., Huang, B., Gamalo-Siebers, M., Wang, J., Yu, G., & Hoaglin, D. C. (2020). The inverse-probability-of-censoring weighting (ipcw) adjusted win ratio statistic: An unbiased estimator in the presence of inde-

pendent censoring. *Journal of biopharmaceutical statistics, 30(5), 882–899*. https://doi.org/https://doi.org/10.1080/10543406.2020.1757692

Dziak, J. J., Li, R., Tan, X., Shiffman, S., & Shiyko, M. P. (2015). Modeling intensive longitudinal data with mixtures of nonparametric trajectories and time-varying effects. *Psychological methods*, *20*(04), 444–469. https://doi.org/https://doi.org/10.1037/met0000048

Efron, B., & Hastie, T. (2016). *Computer age statistical inference : Algorithms, evidence, and data science / bradley efron, stanford university, california, trevor hastie, stanford university, california.* Cambridge University Press.

Elshout, M., Webers, C. A., van der Reis, M. I., de Jong-Hesse, Y., & Schouten, J. S. (2017). Tracing the natural course of visual acuity and quality of life in neovascular age-related macular degeneration: A systematic review and quality of life study. *BMC ophthalmology, 17(1), 120*. https://doi.org/10.1186/s12886-017-0514-3

Faucett, C. L., & Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A gibbs sampling approach. *Statistics in Medicine*, *15*(15), 1663–1685. https://doi.org/https://doi.org/10.1002/(SICI)1097-0258(19960815)15:15⟨1663::AID-SIM294⟩3.0.CO;2-1

Feher, B., Lettner, S., & Heinze, G. (2020). An advanced prediction model for postoperative complications and early implant failure. *Clinical oral implants research*, *31*(10), 928–935. https://doi.org/https://doi.org/10.1111/clr.13636

Fermanian, J.-D., & Scaillet, O. (2004). Some statistical pitfalls in copula modeling for financial applications. *Capital Formation, Governance and Banking*. https://doi.org/10.2139/ssrn.558981

Firth, D. (1993). Bias reduction of maximum-likelihood-estimates. *Biometrika. 1993; 80: 27-38*.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis, 2nd ed.* John Wiley & Sons, Inc. https://doi.org/10.1002/9781119513469

Flaxman, S. R., Bourne, R. R. A., Resnikoff, S., Braithwaite, T., Cicinelli, M. V., Das, A., Jonas, J. B., Keeffe, J., Kempen, J. H., Limburg, H., Naidoo, K., Pesudovs, K., Silvester, A., Stevens, G. A., Tahhan, N., Wong, T. Y., Taylor, H. R., Bourne, R., Ackland, P., … Zheng, Y. (2017). Global causes of blindness and distance vision impairment 1990–2020: A systematic review and meta-analysis. *The Lancet global health*, *5*(12), e1221–e1234.

Frome, E. L. (1983). The analysis of rates using poisson regression models. *Biometrics*, *39*(3), 665–674.

Gelber, R., Gelman, R., & Goldhirsch, A. (1989). A quality-of-life-oriented endpoint for comparing therapies. *Biometrics*, *45*(3), 781–795. https://doi.org/10.2307/2531683

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis, 2nd ed.* Chapman & Hall/CRC.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Ghosal, S., Walker, S. G., Hjort, N. L., Holmes, C., & Müller, P. (2010). The dirichlet process, related priors and posterior asymptotics. In *Bayesian nonparametrics* (pp. 35–79, Vol. Series Number 28). Cambridge University Press.

Gillies, M. C., Campain, A., Barthelmes, D., Simpson, J. M., Arnold, J. J., Guymer, R. H., McAllister, I. L., Essex, R. W., Morlet, N., & Hunyor, A. P. (2015). Long-term outcomes of treatment of neovascular age-related macular degeneration: Data from an observational study. *Ophthalmology (Rochester, Minn.)*, *122*(9), 1837–.

Glymour, M. M., Weuve, J., Berkman, L. F., Kawachi, I., & Robins, J. M. (2005). When is baseline adjustment useful in analyses of change? an example with education and cognitive change. *American Journal of Epidemiology*, *162*(3), 267–278. https://doi.org/10.1093/aje/kwi187

Gómez, G., Plana-Ripoll, O., & Dafni, U. (2016). Selection of the primary end point in an observational cohort study. *Journal of epidemiology and community health (1979)*, *70*(10), 950–953.

Gross, J. G., Glassman, A. R., Jampol, L. M., Inusah, S., Aiello, L. P., Antoszyk, A. N., Baker, C. W., Berger, B. B., Bressler, N. M., Browning, D., Elman, M. J., Ferris, 3., F. L., Friedman, S. M., Marcus, D. M., Melia, M., Stockdale, C. R., Sun, J. K., Beck, R. W., & Network, D. R. C. R. (2015). Panretinal photocoagulation vs intravitreous ranibizumab for proliferative diabetic retinopathy: A randomized clinical trial. *JAMA, 314(20), 2137–2146*. https://doi.org/https://doi.org/10.1001/jama.2015.15217

Größer, J., & Okhrin, O. (2022). Copulae: An overview and recent developments. *WIREs Computational Statistics*, *14*(3), e1557. https://doi.org/https://doi.org/10.1002/wics.1557

Guo, M., & Heitjan, D. (2010). Multiplicity-calibrated bayesian hypothesis tests. *Biostatistics*, *11*(3), 473–483. doi:10.1093/biostatistics/kxq012

Hager, R., Tsiatis, A. A., & Davidian, M. (2018). Optimal two-stage dynamic treatment regimes from a classification perspective with censored survival data. *Biometrics*, *74*(4), 1180–1192. https://doi.org/https://doi.org/10.1111/biom.12894

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning : Data mining, inference, and prediction, second edition / by trevor hastie, robert tibshirani, jerome friedman.* (2nd ed. 2009.). Springer New York.

Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in medicine*, *21*(16), 2409–2419. https://doi.org/10.1002/sim.1047

Hilbe, J. M. (2011). *Negative binomial regression* (2nd ed.). Cambridge University Press.

Hirji, K. F., Tsiatis, A., & Mehta, C. (1989). Median unbiased estimation for binary data. *American Statistician - AMER STATIST*, *43*, 7–11. https://doi.org/10.1080/00031305.1989.10475597

Hirji, K. F., Tan, S. J., & Elashoff, R. M. (1991). A quasi-exact test for comparing two binomial proportions. *Statistics in medicine*, *10(7)*, 1137–1153. https://doi.org/https://doi.org/10.1002/sim.4780100713

Ho, T. W., Ferrari, M. D., Dodick, D. W., Galet, V., Kost, J., Fan, X., Leibensperger, H., Froman, S., Assaid, C., Lines, C., Koppen, H., & Winner, P. K. (2008). Efficacy and tolerability of mk-0974 (telcagepant), a new oral antagonist of calcitonin gene-related peptide receptor, compared with zolmitriptan for acute migraine: A randomised, placebo-controlled, parallel-treatment trial. *Elsevier Ltd*. DOI:%20https://doi.org/10.1016/S0140-6736(08)61626-8

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika 1988; 75:800–802*.

Hoef, J. M. V., & Boveng, P. L. (2007). Quasi-poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology (Durham)*, *88*(11), 2766–2772.

Huang, Y., & Louis, T. (1998). Nonparametric estimation of the joint distribution of survival time and mark variables. *Biometrika*, *85*(4), 785–798. https://doi.org/10.1093/biomet/85.4.785

Hussain, R. M., Neiweem, A. E., Kansara, V., Harris, A., & Ciulla, T. A. (2019). Tie-2/angiopoietin pathway modulation as a therapeutic strategy for retinal disease. *Expert opinion on investigational drugs*, *28*(10), 861–869. https://doi.org/https://doi.org/10.1080/13543784.2019.1667333

Janssen, P., & Veraverbeke, N. (2023). Nonparametric estimation of univariate and bivariate survival functions under right censoring: A survey. *Metrika*, *87*(3), 211–245. https://doi.org/10.1007/s00184-023-00911-7

Jones, A. M., Rice, N., d'Uva, T. B., & Balia, S. (2013). Models for count data. In *Applied health economics* (2nd ed., pp. 295–341). Routledge.

Kalbfleisch, J., & Prentice, R. (2002). The statistical analysis of failure time data. John Wiley; Sons, New York. http://dx.doi.org/10.1002/9781118032985

Kirkwood, J. M., Ibrahim, J. G., Sondak, V. K., Richards, J., Flaherty, L. E., Ernstoff, M. S., Smith, T. J., Rao, U., Steele, M., & Blum, R. H. (2000). High- and low-dose interferon alfa-2b in high-risk melanoma: First analysis of intergroup trial e1690/s9111/c9190 [PMID: 10856105]. *Journal of Clinical Oncology*, *18*(12), 2444–2458. https://doi.org/10.1200/JCO.2000.18.12.2444

Komárek, A. (2024). Mixak 5.8. https://msekce.karlin.mff.cuni.cz/~komarek/

Kosmidis, I., & Firth, D. (2009). Bias reduction in exponential family nonlinear models. *Biometrika. ; 96: 793-804*.

Kruschke, J. K. (2014). *Doing bayesian data analysis: A tutorial with r, jags, and stan, second edition*. Academic Press. 10.1016/B978-0-12-405888-0.09999-2

Kurland, B. F., Johnson, L. L., Egleston, B. L., & Diehr, P. H. (2009). Longitudinal data with follow-up truncated by death: Match the analysis method to research aims. *Statistical science*, *24*(2), 211–222.

Lin, D. Y., & Ying, Z. (1993). A simple nonparametric estimator of the bivariate survival function under univariate censoring. *Biometrika*, *80*(3), 573–581.

Little, T. D. (2024). Longitudinal structural equation modeling. The Guilford Press.

Logxact 12, cytel inc. cambridge, ma. software manual. (2019).

Lovblom, L., Briollais, L., Perkins, B. A., & Tomlinson, G. (2024). Modeling multiple correlated end-organ disease trajectories: A tutorial for multistate and joint models with applications in diabetes complications. *Statistics in Medicine*, *43*(5), 1048–1082. https://doi.org/https://doi.org/10.1002/sim.9984

Lu, K. (2025). Trtswitch 0.1.8. *R package*. https://github.com/kaifenglu/trtswitch

Mallick, H., & Yi, N. (2013). Bayesian methods for high dimensional linear models. *Journal of biometrics & biostatistics*, *1*(005), 163. https://doi.org/https://doi.org/10.4172/2155-6180.S1-005

Mardia, K., Southworth, H., & Taylor, C. (1999). On bias in maximum likelihood estimators. *Journal of statistical planning and inference*, *76*(1), 31–39.

Masic, I. (2022). Medical decision making - an overview. *Acta informatica medica : AIM : journal of the Society for Medical Informatics of Bosnia & Herzegovina : casopis Drustva za medicinsku informatiku BiH*, *30*(3), 230–235. https://doi.org/10.5455/aim.2022.30.230-235

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models / p. mccullagh and j.a. nelder*. (Second edition.). Chapman & Hall/CRC.

Mehta, C. R., Patel, N., & Senchaudhuri, P. (1992). Exact stratified linear rank tests for ordered categorical and binary data. *Journal of Computational and Graphical Statistics*, *1*(1), 21–40. http://www.jstor.org/stable/1390598

Mehta, C. R., & Patel, N. (1995). Exact logistic regression: Theory and examples. *Statistics in medicine*, *14*(19), 2143–2160. https://doi.org/https://doi.org/10.1002/sim.4780141908

Meuli, L., & Dick, F. (2018). Understanding confounding in observational studies. *European Journal of Vascular and Endovascular Surgery*, *55*(5), 737. https://doi.org/https://doi.org/10.1016/j.ejvs.2018.02.028

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, *49*, 65–82.

Mueller, S., & Pearl, J. (2023). *Journal of Causal Inference*, *11*(1), 20220050. https://doi.org/doi:10.1515/jci-2022-0050

Mukhopadhyay, S., & Gelfand, A. E. (1997). Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association*, *92*(438), 633–639.

Nair, A. A., Finn, A. P., & Sternberg Jr, P. (2022). Spotlight on faricimab in the treatment of wet age-related macular degeneration: Design, development and place in therapy. *Drug design, development and therapy*, *16*, 3395–3400.

Nelder, J. A. (1974). Log linear models for contingency tables: A generalization of classical least squares. *Applied Statistics*, *23*(3), 323–329.

Ollila, T., Silvennoinen, J., Joshi, A., Liu, J., Kulathinal, S., & Immonen, I. (2022). Analysing subgroups and treatment discontinuation in a finnish cohort of patients with neovascular amd. *Ophthalmologica. International journal of ophthalmology.*, *245*(4), 358–367. https://doi.org/10.1159/000524848

Ording, A. G., Cronin-Fenton, D., Ehrenstein, V., Lash, T. L., Acquavella, J., Rørth, M., & Sørensen, H. T. (2016). Challenges in translating endpoints from trials to observational cohort studies in oncology. *Clinical epidemiology*, *8*, 195–200.

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. https://doi.org/10.1198/016214508000000337

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press.

Perperoglou, A. (2011). Fitting survival data with penalized poisson regression. *Statistical methods & applications*, *20*(4), 451–462.

Pesarin, F. (1990). On a nonparametric combination method for dependent permutation tests with applications. *Psychotherapy and Psychosomatics, Quality of Life in the Medically III: A Psychosomatic Approach*, *54*(2/3), 172–179. http://www.jstor.com/stable/45113765

Proust-Lima, C. (2024). Lcmm 2.2.0. *R package*. https://cecileproust-lima.github.io/lcmm/

Puhr, R., Heinze, G., Nold, M., Lusa, L., & Geroldinger, A. (2017). Firth's logistic regression with rare events: Accurate effect estimates and predictions? *Statistics in Medicine*, *36*(14), 2302–2317. https://doi.org/https://doi.org/10.1002/sim.7273

Robins, J. M., & Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In J. NP, K. D, & F. VT. (Eds.), *Aids epidemiology: Methodological issues* (pp. 297–331). Birkhäuser Boston. https://doi.org/10.1007/978-1-4757-1229-2_14

Röhmel, J., Gerlinger, C., Benda, N., & Läuter, J. (2006). On testing simultaneously non-inferiority in two multiple primary endpoints and superiority in at least one of them. *Biometrical journal. Biometrische Zeitschrift*, *48*(6), 916–933. https://doi.org/https://doi.org/10.1002/bimj.200510289

Rosenfeld, P. J., Brown, D. M., Heier, J. S., Boyer, D. S., Kaiser, P. K., Chung, C. Y., Kim, R. Y., & Group, M. S. (2006). Ranibizumab for neovascular age-related macular degeneration. *The New England journal of medicine*, *355*(14), 1419–1431. https://doi.org/https://doi.org/10.1056/NEJMoa054481

Satten, G., & Datta, S. (2001). The kaplan–meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician*, *55*(3), 207–210. https://doi.org/10.1198/000313001317098185

Satten, G., Datta, S., & Robins, J. (2001). Estimating the marginal survival function in the presence of time dependent covariates. *Statistics & Probability Letters*, *54*(4), 397–403. https://doi.org/10.1016/S0167-7152(01)00113-4

Scott, J. G., & Berger, J. O. (2006). An exploration of aspects of bayesian multiple testing. *Journal of statistical planning and inference*, *136*(7), 2144–2162.

Seaman, S. R., & White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, *22*(3), 278–295. https://doi.org/https://doi.org/10.1177/0962280210395740

Shah, A. R., & Del Priore, L. V. (2007). Progressive visual loss in subfoveal exudation in age-related macular degeneration: A meta-analysis using lineweaver-burke plots. *American journal of ophthalmology, 143(1), 83–89.* https://doi.org/10.1016/j.ajo.2006.09.043

Sheldon, N. (2019). What does it all mean? *Significance*, *16*(4), 15–17. https://doi.org/https://doi.org/10.1111/j.1740-9713.2019.01296.x

Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika 1986; 63:655–660.*

Smith, G. (2015). 1 - data, data, data. In G. Smith (Ed.), *Essential statistics, regression, and econometrics (second edition)* (Second Edition, pp. 1–25). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-803459-0.00001-7

Smyth, P. (1996). Clustering sequences with hidden markov models. In M. Mozer, M. Jordan, & T. Petsche (Eds.), *Advances in neural information processing systems* (Vol. 9). MIT Press. https://proceedings.neurips.cc/paper_files/paper/1996/file/6a61d423d02a1c56250dc23ae7ff12f3-Paper.pdf

Song, X.-Y., & Lu, Z. (2010). Semiparametric latent variable models with bayesian p-splines. *Journal of Computational and Graphical Statistics - J COMPUT GRAPH STAT*, *19*, 1006210703–19. https://doi.org/10.1198/jcgs.2010.09094

Statxact 12, cytel inc. cambridge, ma. software manual. (2019).

Sullivan, S. G., & Greenland, S. (2013). Bayesian regression in sas software. int j epidemiol. *International journal of epidemiology, 43(3), 974*, *42*, 308–17. https://doi.org/https://doi.org/10.1093/ije/dyt280

Tritchler, D. (1984). An algorithm for exact logistic regression. *Journal of the American Statistical Association*, *79*(387), 709–711. https://doi.org/https://doi.org/10.2307/2288420

Tsiatis, A. A., Degruttola, V., & Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. *Journal of the American Statistical Association*, *90*(429), 27–37. https://doi.org/10.1080/01621459.1995.10476485

Tuuminen, R., Uusitalo-Järvinen, H., Aaltonen, V., Hautala, N., Kaipiainen, S., Laitamäki, N., Ollila, M., Rantanen, J., Välimäki, S., Sipilä, R., Laukkala, T., Komulainen, J., Tommila, P., Immonen, I., Tuulonen, A., & Kaarniranta, K. (2017). The finnish national guideline for diagnosis, treatment and follow-up of patients with wet age-related macular degeneration. *Acta ophthalmologica*, *95*(A105 Suppl), 1–9. https://doi.org/https://doi.org/10.1111/aos.13501

van Buuren, S. (2018). Flexible imputation of missing data, second edition (2nd ed.) Chapman; Hall/CRC. https://doi.org/https://doi.org/10.1201/9780429492259

van Houwelingen, H., H. & Putter. (2011). *Dynamic prediction in clinical survival analysis (1st ed.)*. CRC Press. https://doi.org/https://doi.org/10.1201/b11311

van Smeden, M., de Groot, J. A., Moons, K. G., Collins, G. S., Altman, D. G., Eijkemans, M. J., & Reitsma, J. B. (2016). No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC medical research methodology*, *16*(1), 163. https://doi.org/https://doi.org/10.1186/s12874-016-0267-3

Van Keilegom, I., Akritas, M. G., & Veraverbeke, N. (2001). Estimation of the conditional distribution in regression with censored data: A comparative study. *Computational statistics & data analysis*, *35*(4), 487–500.

Van Keilegom, I., & Veraverbeke, N. (2001). Hazard rate estimation in nonparametric regression with censored data. *Annals of the Institute of Statistical Mathematics*, *53*(4), 730–745. https://doi.org/10.1023/a:1014696717644

Verbeke, G., Fieuws, S., Molenberghs, G., & Davidian, M. (2014). The analysis of multivariate longitudinal data: A review. *Statistical methods in medical research*, *23*(1), 42–59. https://doi.org/10.1177/0962280212445834

Wade, S. (2023). Bayesian cluster analysis. *Phil. Trans. R. Soc., A 381:20220149.* https://doi.org/https://doi.org/10.1098/rsta.2022.0149

Wang, J.-L., & Zhong, Q. (2025). Joint modeling of longitudinal and survival data. *Annual review of statistics and its application*, *12*(1), 449–476.

Wasserstein, R. L., & Lazar, N. A. (2016). The asa statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Westfall, P., & Krishen, A. (2001). Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *Journal of Statistical Planning and Inference; 99:25–41.* https://doi.org/10.1002/pst.291

Wilding, G. E., Chandrasekhar, R., & Hutson, A. D. (2012). A new linear model-based approach for inferences about the mean area under the curve. *Statistics in medicine*, *31*(28), 3563–3578. https://doi.org/https://doi.org/10.1002/sim.5387

Willems, S., Schat, A., & van Noorden, M., M. S. & Fiocco. (2018). Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator. *Statistical methods in medical research*, *27*(2), 323–335. https://doi.org/https://doi.org/10.1177/0962280216628900

Zhao, H., & Tsiatis, A. (1997). A consistent estimator for the distribution of quality adjusted survival time. *Biometrika*, *84*(2), 339–348. https://doi.org/10.1093/biomet/84.2.339

Zhao, H., & Tsiatis, A. (1999). Efficient estimation of the distribution of quality-adjusted survival time. *Biometrics*, *55*(4), 1101–1107. https://doi.org/10.1111/j.0006-341x.1999.01101.x

Zhao, H., Bang, H., Wang, H., & Pfeifer, P. (2007). On the equivalence of some medical cost estimators with censored data. *STATISTICS IN MEDICINE*, *26*, 4520–4530. https://doi.org/10.1002/sim.2882

Ziemer, K. S., Pires, B., Lancaster, V., Keller, S., Orr, M., & Shipp, S. (2018). A new lens on high school dropout: Use of correspondence analysis and the statewide longitudinal data system. *The American statistician*, *72*(2), 191–198.