1 Purpose

The purpose of this research project is to develop and enhance statistical machine learning methods and algorithms for analyzing treatment effects using observational healthcare data characterized by irregular and continuous-time monitoring. The analysis of such data poses significant challenges, and there is currently a lack of statistical methods and tools to account for information being available continuously over time (usually on a daily scale) and the complexity in how doctors' and patients' decisions to initiate, stop, or switch treatment depend on past individual health factors. With a focus on the application in medical and epidemiological research, the project will make methodological statistical advancements to address the inherent challenges and improve the quality and potential of evidence obtained from observational data.

More specifically, the project consists of the following objectives:

Obj. 1: Develop, extend and implement machine learning algorithms to data-adaptively model the continuous-time dependence between treatment decisions, covariate and outcome events and past information on individual health factors.

Obj. 2: Advance the translation between medical questions regarding dynamic administering of treatment and the statistical formulation needed for their estimation, as well as develop and implement efficient and data-adaptive estimators for dynamic treatment effects in continuous-time settings.

Obj. 3: Develop user-friendly software tools implementing continuous-time intensity estimators and targeted learning methods in the statistical software R to make the methods accessible to researchers for analysis of dynamic treatment effects based on observational and experimental longitudinal data, e.g., with irregular monitoring.

1.1 Illustrating applications

To demonstrate the new methods' utility, Obj. 3 will further be devoted to applying the methods to applications where conventional statistical methods fall short, including:

- Analyzing the impact of sustained and discontinued exposure to glucagon like peptide-1 receptor agonist (GLP1-RA) versus metformin on glycemic control, kidney function, and cholesterol levels based on Danish health register data. Unlike current analyses that only focus on treatment initiation, the project's methods allow for proper and flexible adjustment of confounders to accurately balance patient groups who continue or discontinue treatment. This approach is crucial, especially considering the known adverse effects of GLP1 treatment leading to increased treatment discontinuation.

- Evaluating the causal effect of stent surgery administered at a subject-specific time after diagnosis of esophageal cancer on survival among cancer patients using the data from the study Egeland et al. (2022). The original analysis in this study reported hazard ratios from a Cox regression without adjusting for confounders, despite its generally highly limited clinical interpretation and the presence of clear confounding factors that severely unbalance the patients undergoing surgery or not. The results of similarly faulty analyses are now used to guide treatment of patients.

- Handling unbalanced exposure to already approved cardio-protective treatments across treatment arms in placebo-controlled cardiovascular outcomes trials of glucose-lowering treatments. This will be illustrated in the analysis of data from the historical cardiovascular outcomes trial LEADER conducted by Novo Nordisk A/S to assess the cardiovascular risk for liraglutide against placebo, while balancing (time-varying) exposure to concomitant medication such as insulin.

2 Background

Observational healthcare data, such as register data, offer valuable insights into the real-world implementation of evidence-based treatments. These datasets provide detailed information on treatment decisions, disease progression, and individual health factors collected over time, and enables the analysis of dynamic treatment effects in impacting patient survival and disease progression. However, analyzing the effects of treatments administered over time presents challenges that traditional biostatistical methods or standalone machine learning approaches cannot adequately address. For example, comparisons of treatment groups that change over time may be influenced by so-called "immortal time bias", since patients must ultimately survive long enough to switch, initiate or discontinue treatments, potentially introducing bias in the results, and in some cases turning effects upside down (Lange and Keiding, 2014). The complexity of such data really calls for the integration of modern causal inference tools (Robins, 1986; Hernán and Robins). In order to mitigate the methodological challenges inherent in observational research, there has specifically been a growing adoption of the concept of "target trial emulation", as originally proposed by Hernán and Robins (2016). This framework provides a practical approach to making observational studies resemble ideal randomized trials, and ties the notion of hypothetical interventions to a clear definition of effects of (time-varying) treatments or exposures. Modern statistical learning theory (Petersen and van der Laan, 2014) within the targeted learning framework (van der Laan and Rose, 2011, 2018) builds on top of this, emphasizing that not only should observational studies be designed to resemble ideal randomized trials, but the entire statistical analysis plan should be specified beforehand, aligning with the principles of rigorous and transparent research. Despite its simplicity, this way of thinking constitute somewhat of a paradigm shift that has had a profound impact on the way researchers perceive and analyze observational data; combined, these ideas are foundational for the analysis of both experimental and observational data, bridging a gap that has so far been present between advanced statistical theory on the one hand and applied research on the other. More specifically, the targeted learning framework can be effectively combined with machine learning techniques, allowing for data-adaptive modeling of the intricate decision-making processes of doctors and patients, while accurately and efficiently analyzing clinically and scientifically relevant parameters (van der Laan, 2017).

Despite the increasing use of targeted learning, target trial emulation and related causal inference methods, the existing tools are limited to data settings where treatments and outcomes change only at a few fixed time-points (Bang and Robins, 2005; van der Laan and Gruber, 2012). However, in order to answer questions regarding the effects of time-varying treatments in realistic scenarios, the application of these tools requires artificial discretization (Sofrygin et al., 2019). Discretization is evidently related to throwing away data, and has been shown to lead to misleading results and a loss of valuable information (Ferreira Guerra et al., 2020). To overcome these limitations, there is a critical need to develop advanced statistical methods that can effectively handle time-varying treatment effects without discretization in observational healthcare data. While event history analysis has been an active field of (bio)statistical research for many decades, providing researchers within medicine and epidemiology with highly valuable and widely applied tools (Kaplan and Meier, 1958; Cox, 1972; Andersen et al., 1993), these methods have until recently been based on regression modeling approaches for parameters representing associations (by now know to exhibit limited causal interpretation and in most cases failure to answer the questions researchers intend to address, see, Hernán, 2010; Martinussen et al., 2018), and only more recently been approached from a causal inference perspective (Røysland, 2012; Rytgaard et al., 2022; Røysland et al., 2022). Moreover, while flexible machine learning based approaches for survival data has been an active area of research in the cross-field between machine learning and biostatistics (Ishwaran and Kogalur, 2022; Simon et al., 2011; Ozenne et al., 2017), these methods have largely been applied without causal inference tools to provide

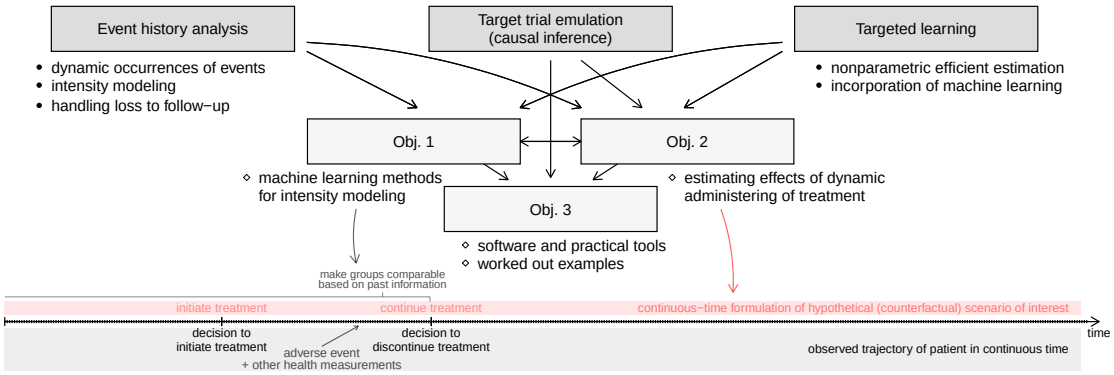

Figure 1: visual overview of the key methodologies and areas of focus.

reliable inference for clinically meaningful parameters.

3 Current state-of-the-art

This research project builds upon foundational tools from efficiency theory (Bickel et al., 1993; van der Vaart, 2000) and targeted learning (van der Laan and Rose, 2011, 2018) to develop new machine learning algorithms that enable statistical inference in continuous-time settings. The project aims to address the limitations of current methods and available implementations, and expand and enhance their application to time-to-event settings. Figure 1 provides a visual overview of the key methodologies and areas of focus. My own work (Rytgaard et al., 2022) has made progress in this area, providing a general theoretical and conceptual framework for targeted learning and data-adaptive causal effect estimation based on a continuous-time counting process framework (Andersen et al., 1993) to handle observations as they happen in time. One challenge in applying these methods is that the formulation of treatment effects must be adapted to the continuous-time setting; particularly, dynamic effects outside the realm of "discrete data" involve manipulating distributions of time-dependent treatment or exposures processes, such as by changing their intensities (Røysland, 2011; Ryalen et al., 2020; Røysland et al., 2022). For this there exist many different choices identifying very different effects and with very different statistical properties. To bridge the gap between theory and practice, and facilitate better practical understanding, the current state-of-the-art methods should be positioned and formulated within a target trial emulation framework. By working out specific examples and applying the methods in practical scenarios, a clearer understanding can be gained, enabling researchers to effectively utilize and interpret the results of their analyses.

The continuous-time targeted learning framework that we propose in Rytgaard et al. (2022) has a clear distinction from other continuous-time causal inference methods proposed, being the first to leverage the power of machine learning techniques and nonparametric efficiency theory for continuous-time causal inference. It requires, however, estimation of conditional expectations entering the martingale integrals forming the efficient influence curve to achieve double robustness and provide the basis for incorporating machine learning procedures, for which no optimal and general solution has yet been proposed. Possible approaches include direct evaluation of the efficient influence curve, or via the highly adaptive lasso along the lines of van der Laan and Rose (2018, Ch. 8). While I have previously worked on specializing and implementing the theoretical advances of Rytgaard et al. (2022) in more classical time-to-event settings (Rytgaard et al., 2023; Rytgaard and van der Laan, 2022), the methods must be extended and focused to be concretely and practically applicable for estimation in large-scale studies with informative timing of treatment. Moreover, as partly discussed in recent work by Røysland et al. (2022), there are intricate subtleties in the definition of intervention on intensities. From a practical

point of view, parameters representing effects under treatment strategies such as 'always-treat' need to be recast in a continuous-time formulation.

To effectively remove confounding, and balance exposure groups over time, machine learning tools should be incorporated to minimize the risk of misspecification and improve precision in the adjustment models. However, whereas a large number of user-friendly machine learning tools are available for regression and classification tasks (Polley et al., 2011, 2021), realistic time-to-event settings must be modeled via continuous-time intensity measures for which the availability of machine learning methods is highly limited. With adaptation of foundational tools from efficiency theory, however, targeted learning estimation procedures are based on the so-called efficient influence curve which reduces the requirements for flexible estimators by exploiting double robustness structures; this has altogether provided the basis for development of new machine learning algorithms that on their own cannot be used to provide statistical inference, but when guided by the efficient influence curve can (van der Laan, 2017). These results are truly foundational for exploiting the power of machine learning in making informed decisions in real-world problems. One approach to build new machine learning algorithms for continuous-time intensities can be accomplished via cross-validation ensembles (Breiman, 1996; van der Laan, 2007) combining recent proposals like Bender et al. (2021) and our own work on the highly adaptive lasso estimator (Rytgaard et al., 2023; Rytgaard and van der Laan, 2022). This has great and immediate practical utility, but involves highly technical work on developing a cross-validation scheme based on likelihood constructions that can combine both continuous and discrete estimators. Particularly, the partial likelihood construction of Cox-based estimators involving the Breslow estimator (Breslow, 1974) for baseline hazards complicates the likelihood construction. In this project, we will proceed from our own practical proposal (Rytgaard and van der Laan, 2022) and that implemented in the widely used software package glmnet (Friedman et al., 2010) to propose a unified and theoretically justified solution. The project will pursue approaches that either extend results by van der Laan and Dudoit (2003), by working with added dominating measures (Kiefer and Wolfowitz, 1956), or by using the method of sieves (Karr, 1987). In either case, extending the estimation to handle time-varying covariates is a further challenge, as it introduces a significantly larger amount of information, making estimation of intensities a high-dimensional problem even if only a few time-varying covariates are used. We aim to provide guidance for estimation of continuous-time intensities in time-varying covariate settings, where choices must be made in terms of what summaries of the history of covariate information to include, but different summaries can be used to form a powerful super learner library (van der Laan, 2007) of learners.

4 Methods

The focus of the work of the project is the development of statistical methods, algorithms and software. We will leverage the strengths of different methodologies and frameworks to address the specific challenges posed by each research objective. A large part of Obj. 1 and 2 consists in combining fundamental frameworks of counting process models and event history analysis (Andersen et al., 1993) and efficiency theory (Bickel et al., 1993; van der Vaart, 2000) with the more modern approaches of targeted machine learning (van der Laan and Rose, 2011, 2018). The general conceptual framework for combining these methodologies is established by Rytgaard et al. (2022), and the project will continue within this framework. To further summarize, some of the existing computational, mathematical, and statistical methods that we will use and extend are:

- Obj. 1: Flexible Poisson-based regression ideas of Bender et al. (2021) and Rytgaard et al. (2023), super learning (Polley et al., 2011, 2021), cross-validation methodology for selection among estimators (van der Laan and Dudoit, 2003), highly adaptive lasso estimation (van der Laan, 2017; Benkeser and van der Laan, 2016);

- Obj. 2: Intensity-based interventions (Røysland, 2011; Ryalen et al., 2020; Røysland et al., 2022), continuous-time targeted learning (Rytgaard et al., 2022);
- Obj. 3: Existing implementations for "discrete-time" targeted learning (Schwab et al., 2014), event history implementions for survival and competing risks analysis (Ozenne et al., 2017).

To ensure the robustness and reliability of the developed methods, we will rigorously evaluate each novel method and compare it to existing methods using simulation studies.

4.1 Software development

In developing the software implementation for Obj. 3, our approach will involve several key steps. First, we will define and outline the desired functionalities of the software, ensuring it can effectively estimate intervention-specific causal parameters and contrasts in general longitudinal settings. To inform our development process, we will conduct a thorough review of existing software implementations for targeting procedures and treatment effect estimation in event history settings. Our development work will be conducted in a modular fashion, with each module serving a specific functionality. This allows us to incrementally build the software, making it more flexible and adaptable to user needs while facilitating continuous improvement and method extensions. By focusing on smaller, specialized modules, we can ensure rigorous testing and optimization before integrating them into the core software. For instance, one module will implement the continuous-time intensity estimators proposed in Rytgaard et al. (2023); Rytgaard and van der Laan (2022). This module will serve as a foundation and can be extended and updated in conjunction with the developments of Obj. 1. Another module will focus on targeting procedures for updating continuous-time intensities in survival and competing risks settings, including an iterative procedure as proposed in Rytgaard et al. (2023) and a one-step procedure for multi-dimensional parameters as proposed in Rytgaard and van der Laan (2023). Additionally, another module will implement estimators for the efficient influence curve, drawing upon the developments of Obj. 2. These modules represent just a few examples.

5 Implementation

Our research implementation will encompass various outputs, including theoretical/methods papers, tutorial papers, and R packages, to ensure the effective utilization of the developed methods. The theoretical/methods papers, targeting esteemed journals such as JASA, Biometrics, Lifetime Data Analysis, and Biostatistics, will serve as primary references for the tools, providing comprehensive theoretical and numerical justifications. To further help increase our visibility in the field, and provide opportunities to meet and connect with new potential collaborators, we will present our work at international conferences.

References

P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer, New York, 1993.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

A. Bender, D. Rügamer, F. Scheipl, and B. Bischl. A general machine learning framework for survival analysis. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III*, pages 158–173. Springer, 2021.

D. Benkeser and M. van der Laan. The highly adaptive lasso estimator. In *Proceedings of the... International Conference on Data Science and Advanced Analytics. IEEE International Conference on Data Science and Advanced Analytics*, volume 2016, page 689. NIH Public Access, 2016.

P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and adaptive inference in semiparametric models*. Johns Hopkins University Press, Baltimore, 1993.

L. Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.

N. Breslow. Covariance analysis of censored survival data. *Biometrics*, 30:89–99, 1974.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

C. Egeland, L. A. Bazancir, N. H. Bui, L. Baeksgaard, J. Gehl, I. Gögenur, and M. Achiam. Palliation of dysphagia in patients with non-curable esophageal cancer–a retrospective danish study from a highly specialized center. *Supportive Care in Cancer*, pages 1–10, 2022.

S. Ferreira Guerra, M. E. Schnitzer, A. Forget, and L. Blais. Impact of discretization of the timeline for longitudinal causal inference methods. *Statistics in medicine*, 39(27):4069–4085, 2020.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

M. A. Hernán. The hazards of hazard ratios. *Epidemiology (Cambridge, Mass.)*, 21(1):13, 2010.

M. A. Hernán and J. M. Robins. Causal inference: what if, 2020. *Boca Raton, FLChapman & Hall/CRC*.

M. A. Hernán and J. M. Robins. Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183(8):758–764, 2016.

H. Ishwaran and U. Kogalur. *Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2022. URL `https://cran.r-project.org/package=randomForestSRC`. R package version 3.1.0.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

A. F. Karr. Maximum likelihood estimation in the multiplicative intensity model via sieves. *The Annals of Statistics*, pages 473–490, 1987.

J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906, 1956.

T. Lange and N. Keiding. Skin cancer as a marker of sun exposure: a case of serious immortality bias. *International journal of epidemiology*, 43(3):971–971, 2014.

T. Martinussen, S. Vansteelandt, and P. K. Andersen. Subtleties in the interpretation of hazard ratios. *arXiv preprint arXiv:1810.09192*, 2018.

B. Ozenne, A. L. Sørensen, T. Scheike, C. Torp-Pedersen, and T. A. Gerds. riskregression: predicting the risk of an event using cox regression models. *The R Journal*, 9(2):440–460, 2017.

M. L. Petersen and M. J. van der Laan. Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology (Cambridge, Mass.)*, 25(3):418, 2014.

E. Polley, E. LeDell, C. Kennedy, and M. van der Laan. *SuperLearner: Super Learner Prediction*, 2021. URL `https://CRAN.R-project.org/package=SuperLearner`. R package version 2.0-28.

E. C. Polley, S. Rose, and M. J. van der Laan. Super learning. In *Targeted Learning*, pages 43–66. Springer, 2011.

J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

K. Røysland. A martingale approach to continuous-time marginal structural models. *Bernoulli*, 17(3):895–915, 2011.

K. Røysland. Counterfactual analyses with graphical models based on local independence. *The Annals of Statistics*, 40(4):2162–2194, 2012.

K. Røysland, P. Ryalen, M. Nygård, and V. Didelez. Graphical criteria for the identification of marginal causal effects in continuous-time survival and event-history analyses. *arXiv preprint arXiv:2202.02311*, 2022.

P. C. Ryalen, M. J. Stensrud, S. Fosså, and K. Røysland. Causal inference in continuous time: an example on prostate cancer therapy. *Biostatistics*, 21(1):172–185, 2020.

H. C. Rytgaard, T. A. Gerds, and M. J. van der Laan. Continuous-time targeted minimum loss-based estimation of intervention-specific mean outcomes. *The Annals of Statistics*, 50(5): 2469–2491, 2022.

H. C. W. Rytgaard and M. J. van der Laan. Targeted maximum likelihood estimation for causal inference in survival and competing risks analysis. *Lifetime Data Analysis*, pages 1–30, 2022.

H. C. W. Rytgaard and M. J. van der Laan. One-step tmle for targeting cause-specific absolute risks and survival curves. *Accepted for Biometika*, 2023.

H. C. W. Rytgaard, F. Eriksson, and M. J. van der Laan. Estimation of time-specific intervention effects on continuously distributed time-to-event outcomes by targeted maximum likelihood estimation. *Accepted for Biometrics*, 2023.

J. Schwab, S. Lendle, M. Petersen, and M. van der Laan. ltmle: Longitudinal targeted maximum likelihood estimation. *R package version 0.9*, 3, 2014.

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 2011. URL `http://www.jstatsoft.org/v39/i05/`.

O. Sofrygin, Z. Zhu, J. A. Schmittdiel, A. S. Adams, R. W. Grant, M. J. van der Laan, and R. Neugebauer. Targeted learning with daily ehr data. *Statistics in medicine*, 38(16):3073–3090, 2019.

J. van der Laan, M. Super learner. *Statistical applications in genetics and molecular biology*, 6 (1):1–23, 2007.

M. J. van der Laan. A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso. *The International Journal of Biostatistics*, 13(2), 2017.

M. J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples, 2003.

M. J. van der Laan and S. Gruber. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The international journal of biostatistics*, 8(1), 2012.

M. J. van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data.* Springer Science & Business Media, 2011.

M. J. van der Laan and S. Rose. *Targeted learning in data science: causal inference for complex longitudinal studies.* Springer, 2018.

A. W. van der Vaart. *Asymptotic statistics.* Cambridge university press, 2000.