Modeling longitudinal epidemiological data using novel methods for statistical learning and regression

Doctoral thesis
to obtain a doctorate (PhD)
from the Faculty of Medicine
of the University of Bonn

Charlotte Caroline Behning
from Dannenberg

2024

Written with authorization of
the Faculty of Medicine of the University of Bonn

First reviewer: Prof. Dr. Matthias Schmid
Second reviewer: Prof. Dr. Inke R. König

Day of oral examination: 03.12.2024

From the Institute of Medical Biometry, Informatics and Epidemiology
Director: Prof. Dr. Matthias Schmid

Table of Contents

List of abbreviations

AIC	Akaike information criterion
AMD	Age-related macular degeneration
CIF	Cumulative incidence function
CNV	Choroidal neovascularization
cRORA	Complete retinal pigment epithelium and outer retinal atrophy
DNN	Deep neural network
GA	Geographic atrophy
iRORA	Incomplete retinal pigment epithelium and outer retinal atrophy
ML	Machine learning
RSF	Random survival forest

1 Abstract

The analysis of longitudinal data plays an important role in medical research. The data is typically collected during follow-up visits in epidemiological observational studies. These studies often investigate the natural history of (slowly) progressing diseases, with endpoints focusing either on changes in outcome variables over time (longitudinal change endpoints) or the time taken to reach a more severe disease stage (time-to-event endpoints).

This dissertation focuses mainly on the application of these methods in ophthalmology based on the experience gained evaluating the MACUSTAR study. The study aims to develop and validate new candidate endpoints for the early stages of age-related macular degeneration (AMD).

This cumulative dissertation consists of four scientific publications that cover several aspects of modeling longitudinal data using novel statistical learning methods and regression, looking into both longitudinal change and time-to-event endpoints. The first project investigates the challenge of recruiting participants with low disease burden. To this end, a Poisson mixed-effects regression model was applied to identify factors associated with increased screening rates of participants with asymptomatic early AMD stages in the multi-center MACUSTAR study. The second work deals with modeling the growth of geographic atrophy (GA) using a novel linear mixed-effects regression framework that directly incorporates the unknown disease age at baseline using random effects. To capture nonlinear GA enlargement, possible transformation parameters were systematically assessed using Box-Cox transformation. The last two publications present approaches to evaluate time-to-event data in the presence of competing events in statistical learning algorithms. Here, an imputation approach was applied, transforming competing event data such that existing single-event methods could be trained. The methods were evaluated using extensive simulation studies and applied on real-world data sets.

All research articles have been accepted for publication in international peer-reviewed journals (see Publications A - D).

2 Introduction

Longitudinal observational data plays an important role in various scientific fields. In a clinical research context, longitudinal data from clinical or epidemiological studies can be used to gain insights into disease progression or to identify underlying risk factors associated with a faster or slower disease progression. Typical endpoints in longitudinal studies may be either defined as a change in a (metric) outcome variable or by time-to-event outcomes.

In the field of ophthalmology, one example would be to study the progression of age-related macular degeneration (AMD), a slowly progressing disease and a leading cause of blindness in older populations (Fleckenstein et al., 2021). The progression of early disease is often symptom-free, but geographic atrophy (GA) and choroidal neovascularisation (CNV), the two late stages of the disease, are associated with severe vision loss (Colijn et al., 2017; Li et al., 2020). To date, no clinical trial endpoints have been developed and accepted by regulators to test new therapies to prevent the progression from early or intermediate AMD stages to late AMD (Terheyden et al., 2021). The development and validation of new candidate endpoints acceptable for clinical trials in intermediate AMD is the main objective of the MACUSTAR study (Finger et al., 2019).

Previous longitudinal AMD studies included both longitudinal change endpoints or time-to-event endpoints. Changes in continuous endpoints included the growth of atrophic lesions in participants with GA (Biarnés et al., 2023), changes in retinal layer thickness (Nittala et al., 2019) or the change of functional outcome measures (Guymer et al., 2014; Terheyden et al., 2021). Examples of time-to-event endpoints in AMD studies are the time to progression to late-stage AMD (Finger et al., 2019), or the time to a visual acuity loss of at least 15 letters (Chew et al., 2014; Terheyden et al., 2021).

In trials employing time-to-event endpoints, early disease-stage participants who are often still symptom-free have to be recruited. Therefore, achieving the required study sample size can be particularly challenging. As part of the MACUSTAR study, we investigated which factors in the management of a multi-center study are crucial to reaching the recruitment goals.

From a statistical perspective, a major challenge in modeling continuous outcomes in late-stage AMD, such as GA growth, arises when the progression over time is assumed to be non-linear and when the disease age at study inclusion is unknown. If participants are enrolled with high variability in GA sizes, modeling of disease progression and identification of risk factors may become problematic. The clinical onset of GA is defined by a small atrophy size of 0.05 mm^2 (Sadda et al., 2018). At this size, vision may not be severely impaired yet, especially if the central fovea is not affected, and often only participants with larger GA sizes are recruited. Therefore, it is relevant to develop suitable statistical methods that can account for different disease ages at the start of the study. Another open problem in modeling GA growth is that there is no consensus in the literature on whether GA enlargement should be assumed linear, quadratic, or exponential (Dreyhaupt et al., 2005; Holz et al., 2007; Feuer et al., 2013; Keenan et al., 2018).

On the other hand, time-to-event endpoints also bear statistical challenges, especially if a disease such as AMD is studied in elderly populations. Here, the study population is not only prone to right-censoring but also exposed to competing events. For example, when the aim is to study progression to late AMD, participants may experience competing events, i.e., death during the study period (Joachim et al., 2015). Additionally, in evaluations in which CNV is regarded as a competing event on the pathway to developing GA (Klein et al., 2008; McGuinness et al., 2021), appropriate methods for competing events must also be applied. If the occurrence of CNV is incorrectly treated as censoring, this type of informative censoring can lead to biased estimates of the cumulative incidence for GA (McGuinness et al., 2021).

While classical regression methods exist that handle competing events, e.g., the subdistribution hazard model by Fine and Gray (Fine and Gray, 1999), only some methods exist for statistical learning approaches (Kantidakis et al., 2023). Statistical learning methods are helpful for developing risk predictions for the development of late AMD from high-dimensional data, e.g., from genetic data or medical imaging. Although, e.g., Ghahramani et al. (2021) and Rivail et al. (2023) applied machine learning (ML) methods to predict progression to late AMD, handling of competing events could not be addressed by the methods used. In these applications, analyses could be greatly simplified if existing and well-established ML methods could be used by shifting the handling of competing events to a preprocessing step.

2.1 Thesis outline: Novel methods for statistical learning and regression

The objective of this cumulative dissertation was to apply and develop novel methods for statistical learning and regression when modeling longitudinal observational data.

While the first article deals with challenges and facilitators when recruiting asymptomatic intermediate AMD participants for longitudinal observational studies (Publication A), the other articles develop and implement novel methods for statistical learning and regression.

Specifically, novel methods were developed for the following situations: (i) in regression analyses, when a non-linear growth rate is assumed (Publication B) and (ii) in statistical learning with time-to-event endpoints in the presence of competing events (Publication C - D).

The appendix lists additional research articles resulting from the work for the MACUSTAR study and consulting projects at the Institute of Medical Biometry, Informatics, and Epidemiology during the years of this PhD.

2.2 Advanced regression

Analyzing disease progression using a continuous outcome variable over time is straightforward when linear progression can be assumed and when the study population has the same disease severity at baseline. However, more advanced modeling strategies are needed if these assumptions do not hold. Also, more advanced regression methods are required if the outcome variable refers to count data.

2.2.1 Count data as the outcome

Count data can occur in longitudinal observational studies, e.g., the occurrence of certain incidents per unit of time. For count data outcomes, standard linear models are inappropriate. Instead, specialized models such as Poisson or negative binomial regression (Hilbe, 2011) can be used to handle the discrete nature of count data (Publication A).

In Publication A, the count data outcome we aimed to model was the number of screenings per week per clinical site in the multi-center, longitudinal, low interventional MACUSTAR study. We applied a mixed-effects model with a Poisson distribution to model factors influencing the screening rates. The covariates of the model included possible recruitment factors and follow-up time since the first site's opened as fixed effect terms. A random effect was used to account for repeated measures per clinical site. As the official start of recruitment was scheduled at different times for each clinical site, the random effect was included as an interaction with the site's activity status.

2.2.2 Non-linear growth rate and varying starting points

If we assume non-linear growth (e.g., of GA size), aligning the timeline to a starting point of the disease is essential. If this cannot be captured appropriately, e.g., by study inclusion criteria, the growth rate might be dependent on the starting point, i.e., the disease age at study entry (Publication B, Fig. 1). In Publication B, we provide a comprehensive framework for modeling the possibly non-linear progression of GA. This involves modeling the unknown disease age directly as a random effect. In addition, a possible non-linear progression is captured via the optimal transformation parameter of a Box-Cox transformation (Box and Cox, 1964), which was determined with regard to the Akaike information criterion (AIC). More details on the modeling strategy can be found in the Method section in Publication B.

2.3 Competing events in statistical learning methods

Analyzing time-to-event endpoints can be complicated by the presence of competing events, which may prevent the occurrence of the events of interest and may violate the non-informative censoring assumption. Handling such competing events in time-to-event statistical learning algorithms is often solved by constructing more complex architectures, e.g., training event-specific sub-networks or event-specific trees, or by using cause-specific splitting rules (Ishwaran et al., 2014; Lee et al., 2018).

In Publication C - D, we used a rather different strategy. Instead of creating new architectures, we transformed the input data in a way that allowed training and estimating the cumulative incidence function (CIF) of the event of interest using single-event ML architectures. The training of single-event architectures offers several advantages. For example, it is less resource-intensive than training multiple cause-specific sub-architectures, which may even be infeasible if only a few competing events are observed. In addition, the statistical properties of single-event architectures are often already better understood and have already been published in simulation studies.

In a longitudinal data set with a time-to-event endpoint, we assume to have observed $i = 1, ..., n$ participants with p baseline covariates $X_i = (x_{i1}, ...x_{ip})^\top$. These participants have either experienced the event of interest (denoted by $\varepsilon_i = 1$), a competing event ($\varepsilon_i = 2$), or remained event-free and were therefore censored. Here, we assume all event times $T_i \in \{1, ..., k\}$ and censoring times $C_i \in \{1, ..., k\}$ on a discrete time-scale, where k denotes the maximum observable time interval.

Analogous to the work of Fine and Gray (1999) and Berger et al. (2020), we are interested in modeling the CIF for the event of interest ($\varepsilon_i = 1$) using a subdistribution hazard approach. The CIF is defined as $F_1(t|X_i) = P(T_i \leq t, \varepsilon_i = 1|X_i)$, or in other words, the probability of experiencing the event of interest at time t or prior given the baseline covariates X_i.

Similar to imputation methods on a continuous time scale proposed by Ruan and Gray (2008), we imputed the unobserved censoring times C_i for all participants who experienced a competing event first. For an illustration of the imputation strategy, see Publication C, Figure 1. The imputation approach uses a life-table estimate of the censoring distribution $G(t) = P(C_i > t)$, estimated on the observed censoring times C_i of the population. For participants who experienced a competing event first, we sampled the unobserved censoring time $\hat{C}_i$ using subdistribution weights. These censoring times $\hat{C}_i$ were then used as imputed censoring times, and a single-event architecture was trained. For further details, see the Method section of Publication C and D.

In Publication C, this imputation approach was evaluated for deep neural networks (DNNs) and in Publication D for random survival forests (RSFs). Publication C compared the performance of the imputation method using both a DNN architecture with event-specific subnetworks (Lee et al., 2018) and a single-event DNN (Ren et al., 2019). In the RSF, several variations of the imputation approach were explored, which included (i) an imputation step before fitting the forest, (ii) imputation in each tree of the forest, and (iii) imputation in each node in all trees.

The performance was evaluated using calibration plots of the CIF, C-index, and integrated Brier score (see Method section of Publication C and D for further details).

2.4 References

Berger M, Schmid M, Welchowski T, Schmitz-Valckenberg S, Beyersmann J. Subdistribution hazard models for competing risks in discrete time. In: Biostatistics 2020; 21 (3): 449–466

Biarnés M, Garrell-Salat X, Gómez-Benlloch A, Guarro M, Londoño G, López E, Ruiz S, Vázquez M, Sararols L. Methodological appraisal of phase 3 clinical trials in geographic atrophy. In: Biomedicines 2023; 11 (6): 1548

Box GE, Cox DR. An analysis of transformations. In: Journal of the Royal Statistical Society: Series B (Methodological) 1964; 26: 211–243

Chew EY, Clemons TE, Agrón E, Sperduto RD, SanGiovanni JP, Davis MD, Ferris FL, Group AREDSR, et al. Ten-year follow-up of age-related macular degeneration in the age-related eye disease study: AREDS report no. 36. In: JAMA ophthalmology 2014; 132 (3): 272–277

Colijn JM, Buitendijk GH, Prokofyeva E, Alves D, Cachulo ML, Khawaja AP, Cougnard-Gregoire A, Merle BM, Korb C, Erke MG, et al. Prevalence of age-related macular degeneration in Europe: the past and the future. In: Ophthalmology 2017; 124 (12): 1753–1763

Dreyhaupt J, Mansmann U, Pritsch M, Dolar-Szczasny J, Bindewald A, Holz F. Modelling the natural history of geographic atrophy in patients with age-related macular degeneration. In: Ophthalmic epidemiology 2005; 12: 353–362

Feuer WJ, Yehoshua Z, Gregori G, Penha FM, Chew EY, Ferris FL, Clemons TE, Lindblad AS, Rosenfeld PJ. Square root transformation of geographic atrophy area measurements to eliminate dependence of growth rates on baseline lesion measurements: a reanalysis of age-related eye disease study report no. 26. In: JAMA ophthalmology 2013; 131: 110–111

Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. In: Journal of the American Statistical Association 1999; 94 (446): 496–509

Finger RP, Schmitz-Valckenberg S, Schmid M, Rubin GS, Dunbar H, Tufail A, Crabb DP, Binns A, Sánchez CI, Margaron P, et al. MACUSTAR: development and clinical validation of functional, structural, and patient-reported endpoints in intermediate age-related macular degeneration. In: Ophthalmologica 2019; 241 (2): 61–72

Fleckenstein M, Keenan TD, Guymer RH, Chakravarthy U, Schmitz-Valckenberg S, Klaver CC, Wong WT, Chew EY. Age-related macular degeneration. In: Nature reviews Disease primers 2021; 7 (1): 31

Ghahramani G, Brendel M, Lin M, Chen Q, Keenan T, Chen K, Chew E, Lu Z, Peng Y, Wang F. Multi-task deep learning-based survival analysis on the prognosis of late AMD using the longitudinal data in AREDS. In: AMIA Annual Symposium Proceedings. Vol. 2021, American Medical Informatics Association, 2021: 506

Guymer RH, Brassington KH, Dimitrov P, Makeyeva G, Plunkett M, Xia W, Chauhan D, Vingrys A, Luu CD. Nanosecond-laser application in intermediate AMD: 12-month results of fundus appearance and macular function. In: Clinical & experimental ophthalmology 2014; 42 (5): 466–479

Hilbe JM. Negative Binomial Regression. Cambridge: Cambridge University Press, 2011

Holz FG, Bindewald-Wittich A, Fleckenstein M, Dreyhaupt J, Scholl HP, Schmitz-Valckenberg S, Group FS, et al. Progression of geographic atrophy and impact of fundus autofluorescence patterns in age-related macular degeneration. In: American journal of ophthalmology 2007; 143: 463–472

Ishwaran H, Gerds TA, Kogalur UB, Moore RD, Gange SJ, Lau BM. Random survival forests for competing risks. In: Biostatistics 2014; 15 (4): 757–773

Joachim N, Mitchell P, Burlutsky G, Kifley A, Wang JJ. The incidence and progression of age-related macular degeneration over 15 years: the Blue Mountains Eye Study. In: Ophthalmology 2015; 122 (12): 2482–2489

Kantidakis G, Putter H, Litière S, Fiocco M. Statistical models versus machine learning for competing risks: development and validation of prognostic models. In: BMC Medical Research Methodology 2023; 23 (1): 51

Keenan TD, Agrón E, Domalpally A, Clemons TE, Asten F van, Wong WT, Danis RG, Sadda S, Rosenfeld PJ, Klein ML, et al. Progression of geographic atrophy in age-related macular degeneration: AREDS2 report number 16. In: Ophthalmology 2018; 125: 1913–1928

Klein ML, Ferris III FL, Armstrong J, Hwang TS, Chew EY, Bressler SB, Chandra SR, Group AR, et al. Retinal precursors and the development of geographic atrophy in age-related macular degeneration. In: Ophthalmology 2008; 115 (6): 1026–1031

Lee C, Zame WR, Yoon J, Schaar M van der. DeepHit: A deep learning approach to survival analysis with competing risks. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018: 2314–2321

Li JQ, Welchowski T, Schmid M, Mauschitz MM, Holz FG, Finger RP. Prevalence and incidence of age-related macular degeneration in Europe: a systematic review and meta-analysis. In: British Journal of Ophthalmology 2020; 104 (8): 1077–1084

McGuinness MB, Kasza J, Wu Z, Guymer RH. Focus on survival analysis for eye research. In: Investigative Ophthalmology & Visual Science 2021; 62 (6): 7–7

Nittala MG, Hogg RE, Luo Y, Velaga SB, Silva R, Alves D, Staurenghi G, Chakravarthy U, Sadda SR. Changes in retinal layer thickness in the contralateral eye of patients with unilateral neovascular age-related macular degeneration. In: Ophthalmology Retina 2019; 3 (2): 112–121

Ren K, Qin J, Zheng L, Yang Z, Zhang W, Qiu L, Yu Y. Deep recurrent survival analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence 2019; 33 (01): 4798–4805

Rivail A, Vogl WD, Riedl S, Grechenig C, Coulibaly LM, Reiter GS, Guymer RH, Wu Z, Schmidt-Erfurth U, Bogunović H. Deep survival modeling of longitudinal retinal OCT volumes for predicting the onset of atrophy in patients with intermediate AMD. In: Biomedical Optics Express 2023; 14 (6): 2449–2464

Ruan PK, Gray RJ. Analyses of cumulative incidence functions via non-parametric multiple imputation. In: Statistics in Medicine 2008; 27 (27): 5709–5724

Sadda SR, Guymer R, Holz FG, Schmitz-Valckenberg S, Curcio CA, Bird AC, Blodi BA, Bottoni F, Chakravarthy U, Chew EY, et al. Consensus definition for atrophy associated with age-related macular degeneration on OCT: classification of atrophy report 3. In: Ophthalmology 2018; 125 (4): 537–548

Terheyden JH, Schmitz-Valckenberg S, Crabb DP, Dunbar H, Luhmann UFO, Behning C, Schmid M, Silva R, Cunha-Vaz J, Tufail A, et al. Use of composite end points in early and intermediate age-related macular degeneration clinical trials: state-of-the-art and future directions. In: Ophthalmologica 2021; 244 (5): 387–395

3 Publications

3.1 Publication A: Challenges, facilitators and barriers to screening study participants in early disease stages-experience from the MACUSTAR study

Terheyden JH, Behning C, Lüning A, Wintergerst L, Basile PG, Tavares D, Melício BA, Leal S, Weissgerber G, Luhmann UFO, Crabb DP, Tufail A, Hoyng C, Berger M, Schmid M, Silva R, Martinho CV, Cunha-Vaz J, Holz FG, Finger RP, on behalf of the MACUSTAR consortium. Challenges, facilitators and barriers to screening study participants in early disease stages - experience from the MACUSTAR study. In: BMC Medical Research Methodology 2021; 21: 1–8

https://doi.org/10.1186/s12874-021-01243-8

Terheyden *et al. BMC Medical Research Methodology* (2021) 21:54
https://doi.org/10.1186/s12874-021-01243-8

BMC Medical Research
Methodology

RESEARCH ARTICLE Open Access

Challenges, facilitators and barriers to screening study participants in early disease stages-experience from the MACUSTAR study

Jan Henrik Terheyden[1*], Charlotte Behning[2], Anna Lüning[1], Ludmila Wintergerst[1], Pier G. Basile[3], Diana Tavares[3], Beatriz A. Melício[3], Sergio Leal[4], George Weissgerber[5], Ulrich F. O. Luhmann[6], David P. Crabb[7], Adnan Tufail[8], Carel Hoyng[9], Moritz Berger[2], Matthias Schmid[2], Rufino Silva[3,10,11], Cecília V. Martinho[3], José Cunha-Vaz[3], Frank G. Holz[1], Robert P. Finger[1*] and on behalf of the MACUSTAR consortium

Abstract

Background: Recruiting asymptomatic participants with early disease stages into studies is challenging and only little is known about facilitators and barriers to screening and recruitment of study participants. Thus we assessed factors associated with screening rates in the MACUSTAR study, a multi-centre, low-interventional cohort study of early stages of age-related macular degeneration (AMD).

Methods: Screening rates per clinical site and per week were compiled and applicable recruitment factors were assigned to respective time periods. A generalized linear mixed-effects model including the most relevant recruitment factors identified via in-depth interviews with study personnel was fitted to the screening data. Only participants with intermediate AMD were considered.

Results: A total of 766 individual screenings within 87 weeks were available for analysis. The mean screening rate was 0.6 ± 0.9 screenings per week among all sites. The participation at investigator teleconferences (relative risk increase 1.466, 95% CI [1.018–2.112]), public holidays (relative risk decrease 0.466, 95% CI [0.367–0.591]) and reaching 80% of the site's recruitment target (relative risk decrease 0.699, 95% CI [0.367–0.591]) were associated with the number of screenings at an individual site level.

Conclusions: Careful planning of screening activities is necessary when recruiting early disease stages in multi-centre observational or low-interventional studies. Conducting teleconferences with local investigators can increase screening rates. When planning recruitment, seasonal and saturation effects at clinical site level need to be taken into account.

Trial registration: ClinicalTrials.gov NCT03349801. Registered on 22 November 2017.

Keywords: Early disease stages, Age-related macular degeneration, Cohort study, Screening, Recruitment

* Correspondence: Jan.Terheyden@ukbonn.de; Robert.Finger@ukbonn.de
[1]Department of Ophthalmology, University Hospital Bonn, Bonn, Germany
Full list of author information is available at the end of the article

Background

Recruiting asymptomatic participants with early disease stages into clinical or epidemiological studies is challenging because these individuals might not be aware of their disease and their perceived disease burden is often low. In order to overcome these challenges, careful planning of screening and recruitment activities is crucial. This includes careful evaluation of screening and recruitment facilitators as well as barriers. A number of studies have reported factors that impact recruitment at different levels [1–7], but knowledge about how to best identify the specific target population of asymptomatic participants with early disease stages into studies remains limited. Against this background we assessed the recruitment process and any measures which impacted screening numbers in a study of early, largely asymptomatic stages of age-related macular degeneration (AMD). Our goal was to retrospectively identify facilitators and barriers to screenings from a sponsor's perspective in a multi-center cohort study of early disease stages.

The reason for addressing early AMD stages in clinical research today is to reduce the signicficant burden of late-stage AMD by developing novel interventions that stop or delay progression from early AMD stages to late AMD and prevent potentially irreversible loss of vision which make late AMD a leading cause of visual loss in industrialised countries [8, 9]. Early stages of AMD progress slowly at an estimated rate of 5–20 per 100 person-years to late AMD [10] and frequently cause no or only little symptoms [11, 12]. Similar to other early disease stages such as early Alzheimer's disease, prediabetes or pre-clinical cancer [13–15], individuals with early stages of AMD are frequently not aware of their disease [16]. This makes it important to investigate which measures facilitate or impede screening activities for clinical studies of early AMD as identified factors are of potential relevance to other studies recruiting asymptomatic participants. We herein report the impact of both facilitators and barriers to screening participants for the MACUSTAR study, a multi-national cohort study focusing on the most high-risk early stage of AMD ("intermediate AMD") from a sponsor's perspective [17].

Methods

The MACUSTAR study

The MACUSTAR study is a multi-centre cohort study focusing mainly on "intermediate AMD", a high-risk type within the early AMD stages. The main study objective is the development of new candidate endpoints for intermediate AMD clinical trials. For this purpose, participants at all AMD disease stages (no, early, intermediate and late AMD) undergo a battery of functional tests and imaging procedures and several patient-reported outcome measures are administered. The majority of participants of the MACUSTAR study has intermediate AMD and was recruited at 20 study sites while the other groups (early AMD, late AMD, no AMD) were recruited only at five study sites. More details on the study protocol including the eligibility criteria, visit schedule, outcome measures and their assessment, confounders, sources of bias and sample size considerations have been published previously [18].

Recruitment for the MACUSTAR study started in March 2018 and lasted for 87 weeks. Patients were screened and recruited at 20 ophthalmological clinical sites in seven European countries (Denmark, France, Germany, Italy, Netherlands, Portugal and United Kingdom). Five of them were academic core partners within the MACUSTAR consortium, the other sites were affiliated with the consortium and members of the European Vision Clinical Research Network (EVICR.net). To facilitate planning of screenings, all sites confirmed their ability to recruit a minimum of 20 individuals into the MACUSTAR study before study initiation; the core partners agreed to a higher target of 40–70 recruited participants. Herein we retrospectively analyse and report the impact of screening measures, and other factors found to be either facilitators or barriers to screening participants.

All institutional ethic committees approved the study and participants gave written informed consent prior to participation. The MACUSTAR project receives funding from the European Union Innovative Medicines Initiative (IMI2) Horizon 2020 programme. It has been registered at the website clinicaltrials.gov with the identifier NCT03349801. Inclusion criteria for this analysis were individuals screened for the MACUSTAR study with the screening diagnosis intermediate AMD, a high-risk type of the early AMD stages (determined at the clinical site). Study inclusion at all study sites was based on the evaluation and confirmation of AMD diagnosis by a central reading centre, as described previously [17]. Exclusion criteria were missing informed consent, participation in any of the other MACUSTAR study groups (early, or late AMD or control group) or relocation to another clinical site within the time of the study. The MACUSTAR clinical study is managed by the academic clinical research organization AIBILI (Association for Innovation and Biomedical Research on Light and Image, www.aibili.pt) and monitored by the European distributed infrastructure network ECRIN-ERIC (www.ecrin.org).

Qualitative evaluation of screening measures

Screening strategies and measures were planned centrally by a coordination team and then implemented through AIBILI and ECRIN-ERIC across all sites. In order to be able to systematically assess the impact of any of these on screening numbers, we extracted all relevant information retrospectively from the study protocol, protocol amendments, clinical site communications such as newsletters and briefings, status reports, meeting minutes and emails from March 2018 to March 2020. All factors that may have contributed to screening rates were compiled and assigned to time periods and clinical sites within the recruitment phase of the study where they had been implemented. Time is measured in weeks since the first site opened. Furthermore, we conducted four in-depth interviews with personnel actively involved in the study (clinical project managers, study site coordinators and research personnel) to identify the most relevant screening factors based on the available screening numbers and factors previously identified. The interviews consisted of two parts to identify additionally relevant factors: Firstly, all interviewed persons were asked to name which factors (a) facilitated screenings or (b) impeded screenings. Secondly, the factors were ranked by the perceived impact on screening numbers by each person interviewed. All persons were interviewed once and only qualitative methods were applied during this step.

Screening data compilation

Due to the availability of devices, ethics approvals, contracting and the necessity to implement the upcoming European Union General Data Protection Regulation in 2018, the first participants were screened at different time points at the participating study sites. After completion of recruitment, data from the electronic case report form and imaging data were collected and cleaned as reported in the study protocol [18]. Screening numbers and recruitment numbers were compiled per clinical site and per week. We assigned week 1 to the first week of recruitment at the first site and used a global consecutive numbering of weeks for all sites. Factors that were considered relevant in the qualitative evaluation (listed chronologically and preceded by a # sign in the results) were assigned to specific clinical sites and to specific time periods within the screening number database based on where and when they were implemented or occurred.

Statistical modelling

We explored the relationship between different screening factors and screening numbers per week at each clinical site. The inter-correlation of recruitment factors was assessed using Pearson correlation coefficients. When two variables correlated with $r > 0.8$, only the factor more strongly associated with screening numbers in qualitative evaluation was used for further analyses. To account for repeated measurements within the study sites, a generalized mixed-effects model was built to investigate the effect of screening factors on the screening numbers per site. As the official start of the recruitment phase was scheduled at different time points for each clinical site, a random intercept depending on the site's activity status was included. Each selected screening factor entered the model as a fixed effect. To capture a possible time-trend, the week number (measured in weeks since first site has opened) was also considered as a fixed linear effect. Since the screening numbers can be treated as count data, we used the Poisson family with a logarithmic link function. Associations between screening numbers and factors contributing to screening numbers are presented in terms of relative risk increases ($\exp(\beta)$) with 95% confidence intervals, where β denotes the coefficient estimate obtained from the mixed-effects Poisson model.

The analyses were performed with the software R, version 3.6.1 (R Core Team 2020, Vienna, Austria), using the packages lme4 and MuMIn [19, 20].

Results

The overall number of screenings for the intermediate AMD group was 767 in 87 weeks. One participant was excluded from the analysis due to relocating after the screening visit. The last site was opened for recruitment 37 weeks after the first site (Fig. 1). The mean screening rate was 0.6 ± 0.9 screenings per week among all sites. At total of 584 participants of the 766 individuals with intermediate AMD included in the analysis (76%) were included in the MACUSTAR study.

Qualitative evaluation

Twenty factors with a possible impact on patient screenings were identified at global study level (Table 1). While some of them occurred continuously, others were linked to specific periods of time.

Screenings in the MACUSTAR study proceeded in three phases. During an *initiation period* (weeks 1–25), the overall screening trend increased and most participating clinical sites were successively opened for recruitment (factor #1). The weekly screening numbers increased noticeably after the summer holiday season of 2018 (#2). Several screening / recruitment measures implemented continuously throughout the MACUSTAR study were initiated in this period, including the recruitment of patients from pre-screening lists, use of dissemination material and a study newsletter as well as individual contacts with investigators (i.e. phone calls and emails to the principal investigator).

In the ensuing *execution period* (weeks 26–60) weekly screening numbers varied more (range: 0–25 total screenings per week). The implemented measures at the beginning of the *execution period* included an increase of participant travel expenses reimbursement from initial

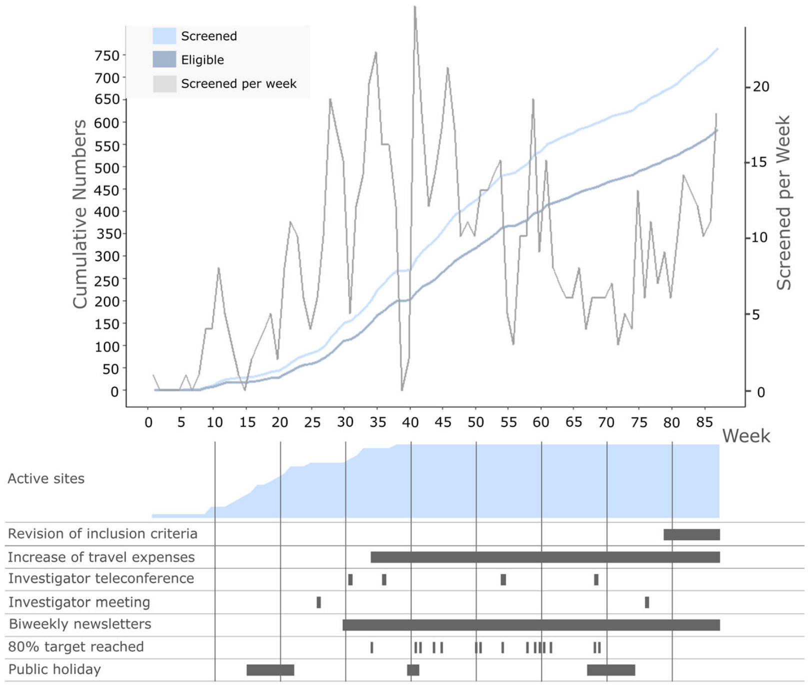

Fig. 1 Participants screened and participants eligible for the MACUSTAR study with intermediate age-related macular degeneration as well as number of active sites and factors impacting screenings, displayed per week since start of recruitment at the first study site (80% target reached refers to individual clinical sites; all other factors are global). The blue curves represent cumulative numbers; the grey curve represents numbers per week

EUR 50 per visit to EUR 75 per visit (#3), an increased MACUSTAR newsletter distribution frequency from monthly to biweekly (#4) and the initiation of regular coordination teleconferences with the project management and monitors (#5). Three teleconferences with the principal investigators (#6) were conducted during the *execution period* and screening numbers increased after these teleconferences. The teleconferences were used to provide data from recent interim analyses to the study staff (principle investigators, study coordinators, study technicians) as well as to allow anyone to ask questions and share approaches on common organizational hurdles, such as the organization of the study schedule, feedback on why screenings failed or pitfalls in the recruitment. Two in person investigator meetings (#7) were also followed by an increase in weekly screenings. The recruitment period was extended beyond the initial end in week 48 until week 87 in order to meet recruitment targets. The two lowest weekly screening rates in the *execution period* (weeks 39 and 56) coincided with Christmas 2018 and the planned end of recruitment before being extended.

The third phase of screenings was a *transition period* (weeks 61–87). It was characterized by more steady screening rates. The number of weekly screenings decreased in the summer holiday season 2019 but increased noticeably afterwards. A change in the inclusion criteria (#8), which opened up recruitment for individuals with unilateral intermediate AMD, was associated with an increase of the screenings at the end of the recruitment period before the transition to the follow-up phase of the study.

At the single clinical site level, the cumulative screenings followed two different patterns. At eight sites, this development increased continuously while at 12 sites, a saturation of the screening rates towards the end of the recruitment period was observed (#9). The core partner sites reached higher recruitment rates (overall median recruitment per site: 66 people) than the other clinical sites (overall median recruitment per site: 29 people; #10).

Variable selection process

Three of the 10 global variables identified (Table 1) were highly correlated (#3 – #5; increase of travel expenses reimbursement, increase of newsletter frequency, initiation of regular coordination teleconferences with the project management and monitors). In the in-depth

Table 1 Relevant global screening factors identified for the MACUSTAR study in qualitative evaluation ordered by estimated magnitude of impact on screening numbers

Screening measures	Factors prioritized in qualitative interviews*	Other factors
All sites	Change of inclusion criteria (opening for individuals with unilateral intermediate disease) (#8)	Dissemination material (patient flyer, referral letter, study procedure flyers, sample visit schedules)
	Increase of participant travel expenses reimbursement (#3)	Distribution of study newsletter
	Investigator teleconferences (#6)	
	Investigator meetings (conferences) (#7)	Letter of appreciation for clinical sites at recruitment start
	Increase of study newsletter frequency (monthly to biweekly) (#4)	Implementation of a clinical site questionnaire to identify unsolved issues
	Regular coordination teleconferences with the project management and monitors (#5)	
Single sites		Pre-screening lists
		Individual contacts with investigators (e-mail, phone, in person)
		Individual contacts with site coordinators
		Appearing in the newsletter as a "top recruiter"
Interacting factors	Public holiday (#2)	Competitive recruitment
	Reaching a high proportion of the initial "recruitment target" or exceeding this target (#9)	Communication of recruiting problems by individual sites
	Successive initiation of screening activity (#1)	Problems with study devices at individual sites
	Consortium core membership (#10)	

* only global factors that could be assigned to specific time periods were allowed
Factors preceded by a # sign were considered relevant in the qualitative evaluation and are displayed in the ranking order obtained in the qualitative evaluation

interviews (see above), higher travel expenses reimbursements were considered to have the largest impact on the overall screening numbers and we therefore included this factor in the multivariable model only. Thus, we identified the following eight parameters for further statistical evaluation in a multivariable model: Modification of inclusion criteria, increase of participant travel expenses reimbursement, organization of investigator teleconferences and meetings in person, public holidays, saturation of screening numbers (80% of overall recruitment per site), week and being a core partner in the MACUSTAR consortium.

Multivariable screening number model

A mixed-effects model including the variables identified qualitatively, excluding highly correlated variables (factors #1 – #3, #6 – #10, Table 1) was fitted to the screening data. The participation at investigator teleconferences, public holidays and reaching a high proportion (80%) of the site recruitment target showed strong associations with screening rates at an individual site level (Table 2). The conditional R^2 value of the model was 0.95 [19, 20]. According to this modelling approach, expected screening numbers increased by the factor exp.(β) = 1.466 (95% CI [1.018–2.112]) after investigator teleconferences were implemented, decreased by the factor exp.(β) = 0.446 (95% CI [0.367–0.591]) during public holidays and decreased with a factor of exp.(β) = 0.669 (95% CI [0.367–0.591]) after a site reached 80% of their recruitment target (after adjusting for the other factors included in the model). This is in line with the average screenings per week, which increased from 0.56 to 0.97 at the time of investigator teleconferences. They decreased from 0.65 to 0.29 during holiday periods and from 0.67 to 0.40 when individual sites reached 80% of their recruitment target.

Discussion

In the MACUSTAR study, we successfully recruited a large cohort of participants with early, mostly asymptomatic AMD stages and found that constant interaction with clinical sites including newsletters, investigator meetings, teleconferences, individual contacts and troubleshooting improve overall recruitment performance. Out of this flurry of activities, however, regular investigator teleconferences were the only measure which was significantly associated with increased screenings at site level. As was to be expected, public holidays were associated with decreased screening performance. Sites slowed down screenings when they reached 80% of their recruitment target. In summary, regular interactions with the site investigators are crucial for a smooth recruitment, and this should likely be increased once sites need to recruit the last 20% as this was when screenings slowed down again.

Table 2 Model parameters for the screening numbers per week in a multivariable generalized mixed-effects model (Poisson family with logarithmic link function)

Predictor	β coefficient*	exp (β)*	95% interval for exp(β)*	*p* value
Revision of inclusion criteria	0.186	1.204	(0.881–1.647)	0.243
Increase of travel expenses reimbursement	0.149	1.161	(0.859–1.570)	0.331
Investigator teleconferences	0.382	1.466	(1.018–2.112)	0.0398
Investigator meetings	−0.084	0.919	(0.705–1.199)	0.534
Core partner site	0.379	1.460	(0.254–8.392)	0.671
Reaching 80% of site recruitment target	−0.357	0.699	(0.542–0.903)	< 0.001
Public holidays	−0.763	0.466	(0.367–0.591)	< 0.001
Week	−0.006	0.994	(0.986–1.003)	0.202
Intercept	−4.19	0.015	(0.005–0.045)	

* adjusted values. No evidence for overdispersion was found (dispersion parameter: 1.0098, $p = 0.3820$ [19]). Only a shared fixed intercept is added when the site is inactive ($\beta_0 = -4.19$, exp.(β_0) = 0.015, 95% CI [0.005–0.045]). Random intercepts α for active clinical sites ranged from 3.42 to 4.35

Failure to recruit a sufficient number of participants in any study can have dire consequences. In an evaluation of two funding agencies, 36% of 195 trials reached less than 80% of their recruitment targets, resulting in a reduced power which had medical, scientific, financial and ethical implications [21, 22]. In addition, low recruitment is a frequent cause for early termination of clinical studies [23]. Careful recruitment planning is therefore an absolute necessity in all clinical studies. In MACUSTAR, no single measure resulted in successful completion of recruitment alone. This is in keeping with available literature where it has previously been noted that only a combination of recruitment measures can lead to successful completion of study recruitment [24].

A review and meta-analysis of recruitment facilitators identified telephone reminders to non-responding candidate participants as a significant facilitator of recruitment to randomized controlled trials [1]. We assume that one of the mediators of this effect was that participants were encouraged to allocate their resources in ways that supported the studies. Similarly, teleconferences with the investigators had a significant positive impact on the MACUSTAR screenings in a multi-site setting. In our experience, teleconferences as well as individual calls allow for a personal relationship and multi- or bidirectional conversations with the site staff on e.g. goals and site-specific difficulties. It also supports peer group learning and creates a common sense of responsibility for the study. In contrast to our findings, Caldwell et al. did not find significantly increased recruitment when keeping increased contact with investigators [2].

Screening rates for the MACUSTAR study dropped significantly during public holidays. This result has not been reported in the available literature [1–3, 5, 21, 25–29] but seems self-evident as facilities are closed during holidays. Gkioni and colleagues reviewed models for the prediction of recruitment when trials are designed [6]. They described that seasonal variations were considered by only 17% of the predictive models found in the literature. We observed high absolute increases in weekly screenings shortly after the end of holiday periods. This finding could be of strategic value for the initiation of recruitment measures in other clinical studies. We assume that facilitating recruitment with new measures could be particularly effective after public holidays.

Besides the assumed influence of teleconferences and public holidays, we observed significant saturation effects of screening numbers in the MACUSTAR study. These would be expected in a study with committed recruitment goals for each clinical site. However, with competitive recruitment in the MACUSTAR study this was an unexpected finding and future research is needed to further assess this effect. In terms of practical implications, sponsor contact should be increased for clinical sites which have almost reached their recruitment target.

Besides these global factors which were present or implemented across all clinical sites in this study, site specific factors such as existing referral networks or a history of clinical research projects are likely to impact screenings numbers and recruitment as well. Unfortunately, it is impossible to assess the impact of any site-specific factors in a systematic fashion as they cannot be quantified across sites.

The main strengths of our analysis include its qualitative and quantitative research methodology, its focus on multi-centre epidemiological research and its inclusion of recruitment factors also identified by previous studies following a thorough review of the literature. We focused our analysis on screening numbers on a site level. Recruitment was not directly assessed in our model since recruited participants out of the pool of screenings were determined by a central reading centre, not by the local investigator. The main limitation of our study is its retrospective character and thus limited generalizability to other studies. As the very few previous studies on recruitment facilitators were done in controlled interventional trials, our results from this observational study have to be interpreted with caution but add to the existing literature. In addition, our analyses provide valuable information in particular relevant to studies recruiting difficult to recruit populations such as early and asymptomatic disease stages [30].

In conclusion, many different facilitators and barriers likely interacted during the recruitment phase of the MACUSTAR study, a multi-site cohort study of early stages of AMD. Regular teleconferences with site investigators increased while public holidays and screening activity saturation at individual clinical sites decreased screening performance. These factors should be given special attention in the design and conduction of future studies as well as selection of clinical sites in particular when recruiting participants with early and largely asymptomatic disease stages.

Abbreviations

AIBILI: Association for Innovation and Biomedical Research on Light and Image; AMD: Age-related macular degeneration; CI: Confidence interval; EVICR.net: European Vision Clinical Research Network

Acknowledgments

We are very grateful for the continuous support provided by Nathalie Seigneuret at the IMI Office in Brussels. We also gratefully acknowledge the work put into this project by consortium members who have since moved on to work elsewhere.

Disclaimer

The communication reflects the author's view and neither IMI nor the European Union, EFPIA, or any associated partners are responsible for any use that may be made of the information contained therein.

Authors' contributions

FGH, RPF, MS, UFOL, DPC, AT, CVM, JCV, RS and CH designed the study. JHT, AL, LW, PGB and DT compiled the dataset. JHT, CB, MB, MS and RPF analysed the data. JHT, CB and RPF wrote the manuscript. All authors contributed substantially to the conception or design of the study, data acquisition, data analysis or data interpretation as well as to drafting the manuscript or critically revising it. They approved the final version to be published.

Funding

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 116076. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

The data proving the main findings of the study are contained within the manuscript. The overall dataset used for analysis is available from the MACUSTAR consortium and the MACUSTAR data access committee upon reasonable request (mail@macustar.eu).

Declarations

Ethics approval and consent to participate

All institutional ethic committees approved the study and participants gave written informed consent prior to participation, as reported previously. These committees included University Hospital Bonn ethics committee (384/17), Paris Ouest IV (04/18_2), AIBILI (032/2017/AIBILI/CE), Nova Medical School (13507/2017), London Queen Square Research Ethics Committee (18/LO/0145), Center for Sundhed Glostrup (H-18000126), Comitato Etico Milano (37910/2018), Ospedale San Raffaele (dated 25/10/2018), Radboudumc technology center (2017–3954) and LUMC commissie medische ethiek (L18.055/SH/sh).

Consent for publication

Not applicable.

Competing interests

J. H. Terheyden: Heidelberg Engineering, Optos, Carl Zeiss MedicTec, CenterVue.
C. Behning: None.
A. Lüning: Heidelberg Engineering, Optos, Carl Zeiss MedicTec, CenterVue.
L. Wintergerst: Heidelberg Engineering, Optos, Carl Zeiss MedicTec, CenterVue.
P. G. Basile: None.
D. Tavares: None.
B. A. Melício: None.
S. Leal: Employee of Bayer AG.
G. Weissgerber: Employee of Novartis Pharma AG.
U. F. O. Luhmann: Employee of F. Hoffmann-La Roche Ltd.
D. P. Crabb: Allergan, Roche, Santen, Centervue.
A. Tufail: None.
C. Hoyng: Optos, Bayer.
M. Berger: None.
M. Schmid: Pixum Vision.
R. Silva: Allergan, Allimera Sciences, Alcon, Bayer, Novartis, Thea.
C. V. Martinho: EVICR.net
J. Cunha-Vaz: Alimera Sciences, Allergan, Bayer, Gene Signal, Novartis, Pfizer, Precision Ocular Ltd., Roche, Sanofi-Aventis, Vifor Pharma and Carl Zeiss Meditec, EVICR.net
F. G. Holz: Acucela, Allergan, Apellis, Bayer, Boehringer-Ingelheim, Bioeq/Formycon, CenterVue, Ellex, Roche/Genentech, Geuder, Grayburg Vision, Heidelberg Engineering, Kanghong, LinBioscience, NightStarX, Novartis, Optos, Pixium, Vision, Oxurion, Stealth BioTherapeutics, Zeiss.
R. P. Finger: Novartis, Bayer, Allergan, Alimera, Roche/Genentech, Santhera, Opthea, Inositec, Ellex, CentreVue, Zeiss, Heidelberg Engineering.

Author details

[1]Department of Ophthalmology, University Hospital Bonn, Bonn, Germany. [2]Institute for Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Bonn, Germany. [3]Association for Innovation and Biomedical Research on Light and Image, Coimbra, Portugal. [4]Bayer AG, Berlin, Germany. [5]Novartis Pharma AG, Basel, Switzerland. [6]Roche Pharmaceutical Research and Early Development, Translational Medicine Ophthalmology, Roche Pharma Research and Early Development, Roche Innovation Center, Basel, Switzerland. [7]Division of Optometry and Visual Sciences, School of Health Sciences, City, University of London, London, UK. [8]Moorfields Eye Hospital, London, UK. [9]Radboud University Medical Center, Nijmegen, Netherlands. [10]University of Coimbra, Coimbra Institute for Clinical and Biomedical Research (iCBR), Faculty of Medicine, Coimbra, Portugal. [11]Ophthalmology Department, Centro Hospitalar e Universitário de Coimbra (CHUC), Coimbra, Portugal.

Received: 4 October 2020 Accepted: 28 February 2021
Published online: 17 March 2021

References

1. Treweek S, Lockhart P, Pitkethly M, Cook JA, Kjeldstrom M, Johansen M, et al. Methods to improve recruitment to randomised controlled trials: Cochrane systematic review and meta-analysis. BMJ Open. 2013;3:e002360. https://doi.org/10.1136/bmjopen-2012-002360 .
2. Caldwell PH, Hamilton S, Tan A, Craig JC. Strategies for increasing recruitment to randomised controlled trials: systematic review. PLoS Med. 2010;7:e1000368. https://doi.org/10.1371/journal.pmed.1000368 .
3. Fletcher B, Gheorghe A, Moore D, Wilson S, Damery S. Improving the recruitment activity of clinicians in randomised controlled trials: a systematic review. BMJ Open. 2012;2:e000496. https://doi.org/10.1136/bmjopen-2011-000496 .
4. Kaur G, Smyth RL, Williamson P. Developing a survey of barriers and facilitators to recruitment in randomized controlled trials. Trials. 2012;13:218. https://doi.org/10.1186/1745-6215-13-218 .

5. Ross S, Grant A, Counsell C, Gillespie W, Russell I, Prescott R. Barriers to participation in randomised controlled trials: a systematic review. J Clin Epidemiol. 1999;52:1143–56.
6. Gkioni E, Rius R, Dodd S, Gamble C. A systematic review describes models for recruitment prediction at the design stage of a clinical trial. J Clin Epidemiol. 2019;115:141–9.
7. Foy R, Parry J, Duggan A, Delaney B, Wilson S, Lewin-Van Den Broek NT, et al. How evidence based are recruitment strategies to randomized controlled trials in primary care? Experience from seven studies. Fam Pract. 2003;20:83–92. https://doi.org/10.1093/fampra/20.1.83 .
8. Bourne RR, Stevens GA, White RA, Smith JL, Flaxman SR, Price H, et al. Causes of vision loss worldwide, 1990-2010: a systematic analysis. Lancet Glob Health. 2013;1:e339–49. https://doi.org/10.1016/S2214-109X(13)70113-X .
9. Colijn JM, Buitendijk GHS, Prokofyeva E, Alves D, Cachulo ML, Khawaja AP, et al. Prevalence of Age-Related Macular Degeneration in Europe: The Past and the Future. Ophthalmology. 2017;124:1753–63. https://doi.org/10.1016/j.ophtha.2017.05.035 .
10. Chakravarthy U, Bailey CC, Scanlon PH, McKibbin M, Khan RS, Mahmood S, et al. Progression from early/intermediate to advanced forms of age-related macular degeneration in a large UK cohort: rates and risk factors. Ophthalmol Retina. 2020;4:662–72.
11. Wu Z, Ayton LN, Luu CD, Guymer RH. Longitudinal changes in microperimetry and low luminance visual acuity in age-related macular degeneration. JAMA Ophthalmol. 2015;133:442–8. https://doi.org/10.1001/jamaophthalmol.2014.5963 .
12. Csaky K, Ferris, F, 3rd, Chew EY, Nair P, Cheetham JK, Duncan JL. Report from the NEI/FDA endpoints workshop on age-related macular degeneration and inherited retinal diseases. Invest. Ophthalmol Vis Sci 2017; 58:3456–3463. doi:https://doi.org/10.1167/iovs.17-22339 .
13. Boada M, Santos-Santos MA, Rodriguez-Gomez O, Alegret M, Canabate P, Lafuente A, et al. Patient engagement: the Fundacio ACE framework for improving recruitment and retention in Alzheimer's disease research. J Alzheimers Dis. 2018;62:1079–90. https://doi.org/10.3233/JAD-170866 .
14. Ebenibo S, Edeoga C, Ammons A, Egbuonu N, Dagogo-Jack S. Pathobiology of Prediabetes in a biracial cohort (POP-ABC) research group. Recruitment strategies and yields for the Pathobiology of Prediabetes in a Biracial Cohort: a prospective natural history study of incident dysglycemia. BMC Med Res Methodol. 2013;13:64. https://doi.org/10.1186/1471-2288-13-64 .
15. das Nair R, Orr KS, Vedhara K, Kendrick D. Exploring recruitment barriers and facilitators in early cancer detection trials: the use of pre-trial focus groups. Trials. 2014;15:98. https://doi.org/10.1186/1745-6215-15-98 .
16. Heinemann M, Welker SG, Li JQ, Wintergerst MWM, Turski GN, Turski CA, et al. Awareness of age-related macular degeneration in community-dwelling elderly persons in Germany. Ophthalmic Epidemiol. 2019;26:238–43. https://doi.org/10.1080/09286586.2019.1597898 .
17. Finger RP, Schmitz-Valckenberg S, Schmid M, Rubin GS, Dunbar H, Tufail A, et al. MACUSTAR: development and clinical validation of functional, structural, and patient-reported endpoints in intermediate age-related macular degeneration. Ophthalmologica. 2018:1–12. https://doi.org/10.1159/000491402 .
18. Terheyden JH, Holz FG, Schmitz-Valckenberg S, Luning A, Schmid M, Rubin GS, et al. Clinical study protocol for a low-interventional study in intermediate age-related macular degeneration developing novel clinical endpoints for interventional clinical trials with a regulatory and patient access intention-MACUSTAR. Trials. 2020;21:659. https://doi.org/10.1186/s13063-020-04595-6 .
19. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. arxiv 2014. doi:https://doi.org/10.18637/jss.v067.i01 .
20. Barton K, Barton MK. Package 'MuMIn'; 2015.
21. McDonald AM, Knight RC, Campbell MK, Entwistle VA, Grant AM, Cook JA, et al. What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. Trials. 2006;7:9.
22. Sully BG, Julious SA, Nicholl J. A reinvestigation of recruitment to randomised, controlled, multicenter trials: a review of trials funded by two UK funding agencies. Trials. 2013;14:166. https://doi.org/10.1186/1745-6215-14-166 .
23. Bernardez-Pereira S, Lopes RD, Carrion MJ, Santucci EV, Soares RM, de OAM, et al. Prevalence, characteristics, and predictors of early termination of cardiovascular clinical trials due to low recruitment: insights from the ClinicalTrials.gov registry. Am Heart J. 2014;168:213–9.e1. https://doi.org/10.1016/j.ahj.2014.04.013 .
24. Gul RB, Ali PA. Clinical trials: the challenge of recruitment and retention of participants. J Clin Nurs. 2010;19:227–33. https://doi.org/10.1111/j.1365-2702.2009.03041.x .
25. Brueton VC, Tierney JF, Stenning S, Meredith S, Harding S, Nazareth I, Rait G. Strategies to improve retention in randomised trials: a Cochrane systematic review and meta-analysis. BMJ Open. 2014;4:e003821. https://doi.org/10.1136/bmjopen-2013-003821 .
26. Watson JM, Torgerson DJ. Increasing recruitment to randomised trials: a review of randomised controlled trials. BMC Med Res Methodol. 2006;6:34.
27. Shaghaghi A, Bhopal RS, Sheikh A. Approaches to recruiting 'Hard-to-Reach' populations into re-search: a review of the literature. Health Promot Perspect. 2011;1:86–94. https://doi.org/10.5681/hpp.2011.009 .
28. Huynh L, Johns B, Liu SH, Vedula SS, Li T, Puhan MA. Cost-effectiveness of health research study participant recruitment strategies: a systematic review. Clin Trials. 2014;11:576–83. https://doi.org/10.1177/1740774514540371 .
29. Ford JG, Howerton MW, Lai GY, Gary TL, Bolen S, Gibbons MC, et al. Barriers to recruiting underrepresented populations to cancer clinical trials: a systematic review. Cancer. 2008;112:228–42. https://doi.org/10.1002/cncr.23157 .
30. Cooper CL, Hind D, Duncan R, Walters S, Lartey A, Lee E, Bradburn M. A rapid review indicated higher recruitment rates in treatment trials than in prevention trials. J Clin Epidemiol. 2015;68:347–54. https://doi.org/10.1016/j.jclinepi.2014.10.007 .

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

3.2 Publication B: Modeling of atrophy size trajectories: variable transformation, prediction and age-of-onset estimation

Behning C, Fleckenstein M, Pfau M, Adrion C, Goerdt L, Lindner M, Schmitz-Valckenberg S, Holz FG, Schmid M. Modeling of atrophy size trajectories: variable transformation, prediction and age-of-onset estimation. In: BMC Medical Research Methodology 2021; 21: 1–12
https://doi.org/10.1186/s12874-021-01356-0

Supplementary information can be found at:
https://doi.org/10.1186/s12874-021-01356-0

Behning *et al. BMC Medical Research Methodology* (2021) 21:170
https://doi.org/10.1186/s12874-021-01356-0

BMC Medical Research Methodology

RESEARCH Open Access

Modeling of atrophy size trajectories: variable transformation, prediction and age-of-onset estimation

Charlotte Behning[1*], Monika Fleckenstein[2], Maximilian Pfau[3], Christine Adrion[4], Lukas Goerdt[5], Moritz Lindner[5], Steffen Schmitz-Valckenberg[2], Frank G Holz[5] and Matthias Schmid[1]

Abstract

Background: To model the progression of geographic atrophy (GA) in patients with age-related macular degeneration (AMD) by building a suitable statistical regression model for GA size measurements obtained from fundus autofluorescence imaging.

Methods: Based on theoretical considerations, we develop a linear mixed-effects model for GA size progression that incorporates covariable-dependent enlargement rates as well as correlations between longitudinally collected GA size measurements. To capture nonlinear progression in a flexible way, we systematically assess Box-Cox transformations with different transformation parameters λ. Model evaluation is performed on data collected for two longitudinal, prospective multi-center cohort studies on GA size progression.

Results: A transformation parameter of $\lambda = 0.45$ yielded the best model fit regarding the Akaike information criterion (AIC). When hypertension and hypercholesterolemia were included as risk factors in the model, they showed an association with progression of GA size. The mean estimated age-of-onset in this model was 67.21 ± 6.49 years.

Conclusions: We provide a comprehensive framework for modeling the course of uni- or bilateral GA size progression in longitudinal observational studies. Specifically, the model allows for age-of-onset estimation, identification of risk factors and prediction of future GA size. A square-root transformation of atrophy size is recommended before model fitting.

Keywords: Geographic atrophy, Age-related macular degeneration, Box-Cox transformation, Mixed-effects models, Prediction, Age-of-onset estimation

Background

Age-related macular degeneration (AMD) is a leading cause of blindness, especially for people in developed countries older than 60 years [1, 2]. AMD has two late stages: choroidal neovascularization (CNV) and geographic atrophy (GA). Here we consider GA, which is thought to be the end stage of AMD when CNV does not develop [3] and which is responsible for vision loss in approximately 20% of all patients with AMD [4]. More than five million people are estimated to be affected by GA worldwide, a number which is supposed to increase with the aging of the population [2]. To date, there is no effective standard treatment available [5].

GA is defined by atrophic lesions of the outer retina resulting from loss of retinal pigment epithelium (RPE), photoreceptors and underlying choriocapillaris (reviewed by [6]). These areas enlarge with time and lead to irreversible loss of visual function [7]. A relevant clinical measure

*Correspondence: charlotte.behning@imbie.uni-bonn.de
[1] Department of Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany
Full list of author information is available at the end of the article

of disease progression is the eye-specific size of GA which can be quantified based on imaging techniques including color fundus photography, spectral domain optical coherence tomography imaging, or fundus autofluorescence (FAF) imaging [8, 9].

A better understanding of the risk factors that accelerate GA size progression is necessary for the development of treatment options, in particular for the design of (interventional) clinical trials. To date, empirical evidence on GA size progression is usually collected through longitudinal observational studies (e.g. [10–12]). In these studies, it is essential to analyze GA size trajectories over time using an adequate statistical model. Specifically, in the absence of a randomized study design, data analysis needs to account for confounding issues as well as correlation patterns, for instance when both eyes of a patient are included in the study. In the latter case, the correlations between the eyes within one patient need to be incorporated as well as the correlations due to repeated measurements over time.

The aim of this analysis is to systematically derive a statistical approach for modeling GA size in observational ophthalmologic studies. As will be demonstrated in the following sections, the proposed approach generalizes various statistical models for GA size progression that have been used in previous publications (see below). Special focus will be given to the following issues, which are considered to be of particular importance for the planning and design of future interventional trials:

(i) Transformation of GA size. Before model fitting, it is important to consider whether the response (here, GA size) should be transformed. Finding an appropriate transformation can provide information about the underlying natural processes that drive the progression of GA. In recent publications on GA size progression, there has been an ongoing discussion about the optimal choice of transformation [11, 13–15]. Three main modeling paradigms have emerged: The first set of models assumes a linear relationship between GA size and covariables (e.g. risk factors or confounding variables). This implies a constant enlargement of GA size over time. Examples of this modeling approach can be found in [13, 14]. The second approach assumes a quadratic enlargement of the lesion size. This is motivated by the thought of circular atrophic lesions that constantly enlarge with their radiuses [11, 15]. The third model type is an exponential model in which atrophic lesions enlarge exponentially. Compared to a linear growth model, Dreyhaupt et al. [13] found that the assumption of exponential growth led to improved model fits.

(ii) Age-of-onset estimation. Another relevant topic for modeling GA size progression is the estimation of the age of disease onset. Research on this topic is motivated by the fact that in many clinical trials patients can only be included when the disease is already manifested in a later stage. The estimated age-of-onset may, in contrast to lesion size, be considered as time-invariant variable, and facilitate association analyses with other time-invariant variables such as the genotype.

(iii) Identification of risk factors and confounding variables. For the development of AMD treatments, it is essential to specify meaningful inclusion and exclusion criteria for use in future clinical trials. It is therefore of high importance to identify relevant risk factors and confounding variables, and to analyze their relationships with GA size progression. Such an analysis can be achieved by building a multivariable regression model from observational data that includes relevant risk factors and confounders as covariables.

To address the issues described above, we derive a statistical regression model that includes (possibly transformed versions of) GA size as response variable, as well as potential risk factors and/or confounders (such as e.g. age, smoking) as covariables. To account for the above mentioned correlations between eyes of the same patient as well as temporal correlations, we investigate the use of a mixed-effects modeling approach with patient- and eye-specific random effects terms. In this framework, we identify the "optimal" transformation of GA size by conducting a systematic search within the family of Box-Cox transformations [16]. As will be shown, this systematic approach also allows for the derivation of formulas for age-of-onset estimation. Furthermore, we demonstrate how predictions of future (untransformed) GA size values can be obtained from the fitted regression model.

For model derivation and illustration, we will apply the proposed methods to a data set collected by the multicenter *Fundus Autofluorescence in AMD* (FAM) study (NCT00393692) and by its single-center extension study, the *Directional Spread in Geographic Atrophy* (DSGA) study (NCT02051998). These noninterventional, prospective natural history studies adhered to the tenets of the Declaration of Helsinki and were approved by the institutional review boards of the participating centers. Written informed consent was obtained from each participant after explanation of the studies' nature and possible consequences of participation.

Methods

Data

The data set used here was collected from patients with GA secondary to AMD that were recruited for the FAM study and followed-up in the DSGA study.

The inclusion and exclusion criteria have been described elsewhere [14, 17]. In brief, the two studies included eyes without any history of retinal surgery, radiation therapy, laser photocoagulation or retinal diseases other than AMD. GA size measurements were obtained by grading FAF retinal images that were recorded at the baseline and follow-up visits. Data was only used for statistical analysis if the difference in total GA size between two graders was smaller than 0.15 mm^2 and if the patients had at least two visits.

Our analysis data set contained $N = 150$ eyes from $n = 101$ patients that where examined in up to nine follow-up visits. At baseline, the median age was 75.7 years (IQR: $70.7 - 80.6$ years); 61.4% of the patients were female, and the mean follow-up time was 3.36 years (range $0.5 - 13.7$ years) due to the extension by the second study. The GA size varied strongly between eyes: mean GA size at baseline was 5.64 mm^2, ranging between 0.07 mm^2 and 31.41 mm^2. The status of hypertension and hypercholesterolemia was assessed by a patient-reported questionnaire at the baseline visit. Information was obtained based on patients' reports and current medication; medical reports were included in the assessment if available. For details see Table 1.

Regression modeling

Within a typical ophthalmologic study setting, patients participate in several follow-up visits at which one or both eyes are examined. This leads to correlated measurements, both within the patients and over time. Thus, a model is needed that captures complex correlation structures. A popular regression model, which has been used regularly in the literature on GA [11, 13] and which is also considered here, is a mixed-effects model with random effects terms for both eye and patient. Yet, there exists a variety of model specifications and the specific structure is still a matter of debate [18].

Before introducing the full mixed-effects model with possible risk factors and confounders, we start with a model that contains a time trend as only (continuous) covariable. This model serves as a basic model that captures the time dependency of GA enlargement.

Table 1 Characteristics of the analysis data set used for statistical modeling

	Count	Percent
Patients (n)	101	
Eyes (N)	150	
Bilateral GA	49	48.50%
Unilateral GA	52	51.50%
Hypertension		
yes	56	55.40%
no	44	38.60%
Hypercholesterolemia		
yes	28	27.70%
no	70	69.30%
No. of patients with no. of visits		
2 visits	25	24.75%
3 visits	23	22.78%
4 visits	23	22.78%
5-9 visits	30	29.70%
	Mean (Range)	Median (IQR)
Age at baseline	75.61	75.66
[years]	(57.23 - 95.06)	(70.67 - 80.62)
Follow-up time	3.36	2.90
[years]	(0.50 - 13.70)	(1.61 - 4.57)
GA size at baseline	5.64	4.30
mm^2	(0.07 - 31.40)	(1.76 - 7.60)

All data considered in this paper was collected from patients with GA secondary to AMD that were recruited for the FAM study. If further monitoring of these patients was performed via the DSGA study, the further progression is included in the analysis data set

Mixed-effects model with time as only covariable. As suggested by Shen et al. [18], we follow the hypothesis that the progression of GA has an underlying process of GA expansion that is mostly the same over time for all eyes. Differences in eyes may arise due to different exposition to environmental conditions, and, most importantly, GA size varies between patients as they enter the study at different time points in their disease history. We therefore propose to include the disease age $\Delta_i \geq 0$ of an eye i at study entry directly in the model. We further assume that the atrophy size y_{it} of an eye i depends on the (unknown) age of the disease at study entry Δ_i and the (observable) follow-up time $t \geq 0$ that has passed since. Time is assumed to be measured on a continuous scale, e.g. in days or years since baseline. Under the assumptions by Shen et al. [18], and considering (for the moment) a linear enlargement of GA, this leads to the following regression model:

$$y_{it} = \beta \cdot (\Delta_i + t) + \epsilon_{it}\,, \tag{1}$$

where β denotes the regression slope (i.e. the constant enlargement rate). The residuals ϵ_{it}, $i = 1, \ldots, N$, are assumed to be normally distributed with zero mean and variance σ^2.

If it is further assumed that the disease age at study entry can be approximated by a normal distribution, the model in (1) can be parameterized such that it becomes a linear mixed-effects model. This is seen by defining $\theta_i := \beta \cdot \Delta_i \sim \mathcal{N}\left(\mu_\theta, \sigma_\theta^2\right)$ and $\alpha_i := \theta_i - \mu_\theta \sim \mathcal{N}\left(0, \sigma_\theta^2\right)$, so that Model (1) can be written as

$$y_{it} = \mu_\theta + \beta t + \alpha_i + \epsilon_{it}. \tag{2}$$

In this form, the model reads as follows: The atrophy size y_{it} depends on a fixed intercept μ_θ, an eye-specific random intercept α_i that reflects the deviation of the disease age of eye i at study entry from the mean disease age at study entry, and an overall linear time trend βt that is the same for all eyes.

When there are patients in the study that contributed data from both eyes, one needs to consider the nested data structure and account for the correlations between measurements taken from the same patient. This can be done by extending the model equation as follows:

$$y_{ijt} = \mu_\theta + \beta t + \zeta_j + \alpha_i + \epsilon_{ijt}, \tag{3}$$

where $\zeta_j \sim \mathcal{N}\left(0, \sigma_\zeta^2\right)$, $j = 1, \ldots, n$, is a normally distributed patient effect and α_i the effect of an 'eye within a patient'. Note: While it is assumed that the residual terms ϵ_{ijt} are independent of the random effects α_i and ζ_j, the latter two terms are generally allowed to be correlated. For simplicity, and without loss of generality, we will assume independence of all random effects terms in the following.

Mixed-effects model with covariables. When introducing covariables into the model, it is reasonable to assume that risk factors and/or confounders equally influence the enlargement of GA before and after inclusion of an eye in the study. This assumption can be incorporated in Model (1) by adding a covariable-dependent slope to the model equation:

$$y_{it} = \left(\beta + \boldsymbol{\beta}_x^\top \boldsymbol{x}_i\right) \cdot (\Delta_i + t) + \epsilon_{it}\,, \tag{4}$$

where $\boldsymbol{x}_i = (x_1, ..., x_k)_i^\top$ is a vector of k (possibly time-dependent) risk factors for each eye and $\boldsymbol{\beta}_x = \left(\beta_{x_1}, ..., \beta_{x_k}\right)^\top$ is a vector of parameters that accelerate or slow down GA size progression ($\beta_{x_s} > 0$ and $\beta_{x_s} < 0$, respectively, $s \in \{1, \ldots k\}$). Note that in the following, we will not distinguish between risk factors and confounders any more, as we assume that both are collected in the vectors $\boldsymbol{x}_i$.

Similar to the reparametrization used above, we write $\Delta_i := (\mu_\Delta + \gamma_i) \sim \mathcal{N}\left(\mu_\Delta, \sigma_\Delta^2\right)$, where μ_Δ and σ_Δ^2 denote the mean and the variance of the i-the eye at study entry.

The mixed-effects model with covariables can then be written as

$$\begin{aligned} y_{it} &= \left(\beta + \boldsymbol{\beta}_x^\top \boldsymbol{x}_i\right)\, \mu_\Delta + \left(\beta + \boldsymbol{\beta}_x^\top \boldsymbol{x}_i\right)\, \gamma_i + \beta t + \boldsymbol{\beta}_x^\top \boldsymbol{x}_i t + \epsilon_{it} \\ &= \beta \mu_\Delta + \beta t + \mu_\Delta \boldsymbol{\beta}_x^\top \boldsymbol{x}_i + \boldsymbol{\beta}_x^\top \boldsymbol{x}_i t + \beta \gamma_i + \boldsymbol{\beta}_x^\top \boldsymbol{x}_i \gamma_i + \epsilon_{it}. \end{aligned} \tag{5}$$

with eye-specific random effects $\gamma_i \sim \mathcal{N}\left(0, \sigma_\Delta^2\right)$. The linear enlargement in Model (5) thus implies dependency of y_{it} on an interaction term between t and $\boldsymbol{x}_i$, and also on random slopes of the covariable values $\boldsymbol{x}_i$. Importantly, Eq. 5 implies numerous dependencies between the slope parameters associated with t, $\boldsymbol{x}_i$, $\boldsymbol{x}_i t$, γ_i, and $\boldsymbol{x}_i \gamma_i$, so that the model no longer possesses the structure of a "standard" mixed-effects model with unrestricted estimation of coefficients. Details on model fitting will be given below.

Finally, when considering patients that contribute data from both eyes, one specifies

$$\begin{aligned} y_{ijt} = {} & \beta \mu_\Delta + \beta t + \mu_\Delta \boldsymbol{\beta}_x^\top \boldsymbol{x}_i + \boldsymbol{\beta}_x^\top \boldsymbol{x}_i t + \beta \gamma_i \\ & + \boldsymbol{\beta}_x^\top \boldsymbol{x}_i \gamma_i + \beta \zeta_j + \boldsymbol{\beta}_x^\top \boldsymbol{x}_i \zeta_j + \epsilon_{ijt} \end{aligned} \tag{6}$$

with patient-specific random effects $\zeta_j \sim \mathcal{N}\left(0, \sigma_\zeta^2\right)$, $j = 1, \ldots, n$, and an additional interaction term between $\boldsymbol{x}_i$ and ζ_j.

The model equations presented so far ascribe a linear relationship between time, risk factors, and GA size. In the following section, possible transformations are examined, so that the modeling approach is extended to model nonlinear progressions.

Transformation of the response

As an example, Fig. 1 A shows the GA size trajectories of four eyes contained in the analysis data set. Considering

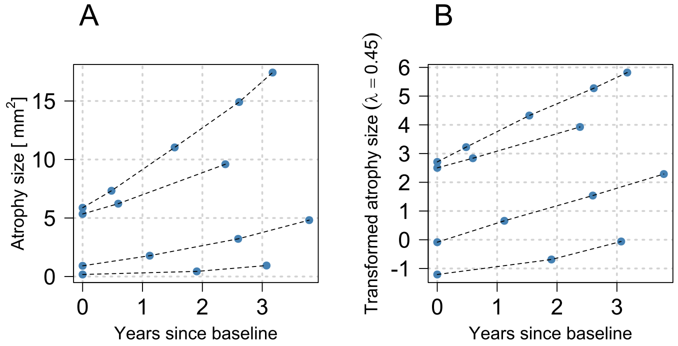

Fig. 1 Progression of GA size. **A** Untransformed GA size trajectories of four different eyes contained in the analysis data set and **B** trajectories on a transformed scale with transformation parameter λ = 0.45

these progressions, it is conceivable to assume that the trajectories are not strictly linear. Since the model equations above (Models (1) to (6)) refer to linear enlargement processes, a transformation of the response is convenient for modeling non-linear progression (see Fig. 1B).

Three different transformation approaches have been used in recent publications on GA size progression (e.g. [11, 13–15]): (i) Linear models with no response transformation implying a linear relationship between GA size and the covariables, (ii) linear models with square root transformation of the response, and (iii) linear models with log-transformed response – or equivalently exponentially transformed models with no transformed response – implying an exponential enlargement of the lesion size.

Box-Cox transformation Instead of comparing only the most commonly used transformations, we consider a systematic and more comprehensive strategy for finding an appropriate transformation of the GA size. For this systematic approach, the Box-Cox model class is applied because it covers a wide range of transformations, including the transformations (i) to (iii) above. More specifically, for an atrophy size $y > 0$ we consider the class of Box-Cox transformations

$$f_\lambda(y) := y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \log(y) & \text{if } \lambda = 0, \end{cases} \quad (7)$$

as introduced by [16]. Applying (7) to one of the Models (1)-(6) reads as follows: $\lambda = 1$ refers to a model with no response transformation, $\lambda = 0.5$ corresponds to a square-root transformation of the response and $\lambda = 0$ can be interpreted as exponential enlargement of the GA size.

Model comparison The main criterion used for our model comparisons was Akaike's Information Criterion (AIC) [19]. More specifically, our aim was to choose the transformation parameter λ that minimized AIC on the analysis data set while assuring that the assumptions of Models (1) to (6) were best possibly met, in particular the normality of the residuals. The AIC is defined by $\text{AIC} = -2 \cdot \log(L) + 2 \cdot n_{\text{params}}$, where L is the likelihood of the model under consideration (evaluated at the maximum likelihood estimate) and n_{params} denotes the number of parameters used in the model. As we compared models with a transformed response, we applied the density transformation theorem to compute the likelihood L.

Maximum likelihood estimation The estimation of the model parameters was performed by maximum likelihood (ML) estimation. ML estimation was carried out for a grid of fixed transformation parameters λ using the transformed GA size values. Subsequently, the likelihoods were compared and the transformation parameter referring to the model with minimum AIC was considered best.

We initially assumed that there was an "optimal" value λ for which the transformed atrophy size given the random effects followed a normal distribution. In addition, we briefly considered random effects with an unspecified mixing distribution as a non-parametric cross-check. The two approaches will be described in the next paragraphs.

Normally distributed random effects As noted above, the linear model in (6) imposes numerous side conditions on the slope parameters associated with t, $\boldsymbol{x}_i$, $\boldsymbol{x}_i t$, γ_i, and $\boldsymbol{x}_i\gamma_i$. In order to fit Model (6) using readily available software for the estimation of the slope parameters (without side conditions, such as the R add-on package **lme4**[20], version 1.1-25), we propose to iterate the following steps:

(i) For given estimates $\hat{\beta}$ and $\hat{\boldsymbol{\beta}}_x$ compute the values of the working covariable $\tilde{x}_i := \hat{\beta} + \hat{\boldsymbol{\beta}}_x^\mathsf{T} \boldsymbol{x}_i$.

(ii) Fit the linear mixed-effects model

$$y_{ijt} = \beta t + \boldsymbol{\beta}_x^\mathsf{T} \boldsymbol{x}_i t + \mu_\Delta \tilde{x}_i + \tilde{x}_i \gamma_i + \tilde{x}_i \zeta_j + \epsilon_{ijt} \quad (8)$$

to obtain updates of the coefficient estimates of $\hat{\mu}_\Delta$, $\hat{\beta}$, and $\hat{\boldsymbol{\beta}}_x$. Note, that Model (8) is just a re-formulation of Model (6) that can be fitted without side conditions on its slope parameters. For the fitting procedure a fixed intercept term is added to increase computational stability and to relax the condition that the empirical mean of estimated random effects terms is forced to be zero.

The starting values for $\hat{\beta}$ and $\hat{\boldsymbol{\beta}}_x$ in Step (i) may be obtained from (8) with an initial value of $\tilde{x}_i = 1$. As demonstrated in the supplementary materials (see Additional file 1), repeated execution of (i) and (ii) will typically converge to the final estimates after less than 20 iterations.

Random effects with unspecified mixing distribution As an alternative to mixed-effects modeling with normally distributed terms, Almohaimeed et al. [21] proposed to consider a nonparametric maximum likelihood (NPML) approach. This approach approximates the distribution of each random effect by a discrete distribution with finite number of mass points K. It then uses an expectation-maximization algorithm to find the nonparametric maximum likelihood estimate. Here, the NPML approach is used to verify the optimal transformation parameter obtained from modeling with normally distributed random effects.

Age-of-onset estimation

Model without covariables As defined by [22], a diagnosis for GA can be given at a minimum lesion diameter of 250 μm and thus a lesion area of 0.05 mm^2. Based on this specification and denoting λ_{opt} as the value of λ that is optimal w.r.t. AIC, the time $\hat{t}_{0_{ij}}$ at which the atrophy size was $\hat{y}_{ijt_0} = 0.05[\,\text{mm}^2]$ (i.e. $\hat{y}^{(\lambda)}_{ijt_0} = \lambda^{-1}_{opt} \cdot (0.05^{\lambda_{opt}} - 1)$) can be obtained by solving the model equation of the transformed mixed-effects Model (3) for t:

$$\hat{t}_{0_{ij}} = \frac{\lambda^{-1}_{opt} \cdot \left(0.05^{\lambda_{opt}} - 1\right) - \left(\hat{\mu}_\theta + \hat{\zeta}_j + \hat{\alpha}_i\right)}{\hat{\beta}}, \qquad (9)$$

where $\hat{\beta}$ and $\hat{\mu}_\theta$ denote the ML estimates of β and μ_θ, respectively, and $\hat{\zeta}_j$ and $\hat{\alpha}_i$ denote the realizations of the random effect terms. As a consequence, subtracting the estimated time $\hat{t}_{0_{ij}}$ from the patient's age at study entry results in the estimated age-of-onset of GA in the i-th eye of patient j. Remark: While from a modeling perspective a theoretical atrophy size of $y_{ijt_0} = 0\,\text{mm}^2$ could be defined at the time of disease onset, we will focus on the clinically relevant definition $\left(y_{ijt_0} = 0.05\,\text{mm}^2\right)$ here. For $y = 0$ it holds that $t_{0_{ij}} = \Delta_{ij} = \frac{1}{\beta} \cdot (\mu_\theta + \zeta_j + \alpha_i)$.

Model with covariables Analogous to (9) one can estimate the ages of GA onset of the study eyes in a model with additional covariables. From Eq. 8 one obtains

$$\hat{t}_{0_{ij}} = \frac{\lambda^{-1}_{opt} \cdot \left(0.05^{\lambda_{opt}} - 1\right) - \tilde{x}_i\left(\hat{\mu}_\Delta + \hat{\zeta}_j + \hat{\alpha}_i\right)}{\tilde{x}_i} \qquad (10)$$

where $\tilde{x}_i := \hat{\beta} + \hat{\boldsymbol{\beta}}^{\mathsf{T}}_x \boldsymbol{x}_i$ contains the parameters obtained from ML estimation.

Prediction

Evaluating a model and its coefficients only on a transformed scale is challenging as the linearity of the predictor-response relationships in Models (5) and (6) only holds on the transformed scale but not on the original scale of the response (provided that $\lambda \neq 1$). As a consequence, the calculation of the expected GA size $\mathbb{E}(y|\boldsymbol{x})$ – and hence any prediction of expected disease progression – cannot be done in an unbiased way by a simple back-transformation.

To see this, consider a non-linear Box-Cox transformation $f(y)$ with an arbitrary parameter $\lambda \neq 1$ and, where existent, the corresponding inverse Box-Cox transformation $f^{-1}(y)$. Further, let $f(y_{ijt}|\boldsymbol{x}_i) = z_{ijt} + \epsilon_{ijt}$, where $z_{ijt} := \mathbb{E}(f(y_{ijt}|\boldsymbol{x}_i))$ and ϵ_{ijt} denote the linear predictor and the residual, respectively in one of the above models. A naive back-transformation would directly take the inverse of the linear predictor, i.e. $f^{-1}(z_{ijt})$, which differs from the desired expected GA size value $\mathbb{E}(y_{ijt}|\boldsymbol{x}_i) = \mathbb{E}\left(f^{-1}\left(z_{ijt} + \epsilon_{ijt}\right)\right)$ by Jensens's inequality [23]. In other words, $f^{-1}\left(\mathbb{E}\left(f\left(y_{ijt}|\boldsymbol{x}_i\right)\right)\right) \neq \mathbb{E}(y_{ijt}|\boldsymbol{x}_i)$. To address this issue and to obtain unbiased predictions of the GA size, we propose to sample $r = 10,000$ residuals from the empirical distribution $\hat{\epsilon}_1, ..., \hat{\epsilon}_r$ in the respective fitted model. The expected atrophy size on the original scale can then be estimated by $\widehat{\mathbb{E}(y_{ijt}|\boldsymbol{x}_i)} := \frac{1}{r}\sum_{u=1}^{r} f^{-1}\left(\hat{z}_{ijt} + \hat{\epsilon}_u\right)$, where $\hat{z}_{ijt}$ denotes the fitted value of $f(y_{ijt}|\boldsymbol{x}_i)$.

Results

In this section, we present the results obtained from fitting Models (2), (3) and (6) to the analysis data set (150 eyes of 101 patients). Missing values in the covariables were imputed using the R package **mice** [24] with one imputation run. Fitting was done using **lme4** [20] with the algorithm described above.

Modeling of GA size trajectories

Determination of the transformation parameter In order to determine the optimal value of the transformation parameter λ, we evaluated linear mixed-effects models of the forms (3) and (6) on the analysis data set. Box-Cox-transformed responses with varying values of λ were considered in each of the models. As seen in Fig. 2A, the minimum AIC value was reached at $\lambda_{opt} = 0.45$ in the model without covariables. The model with covariables

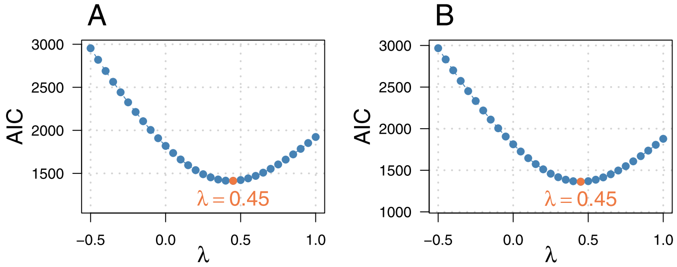

Fig. 2 Determination of the optimal Box-Cox transformation. For each value of the transformation parameter λ, parametric mixed-effects models **A** without covariables as in Model (3) and **B** with covariables as in Model (6) were fitted to the analysis data set. Model fitting was performed using the R package **lme4**. The orange dot indicates the optimal fit, which was achieved at $\lambda_{opt} = 0.45$. For Model (3), the optimal AIC value was $AIC_{\lambda=0.45} = 1413.69$ and for Model (6) the optimal AIC value was $AIC_{\lambda=0.45} = 1347.78$

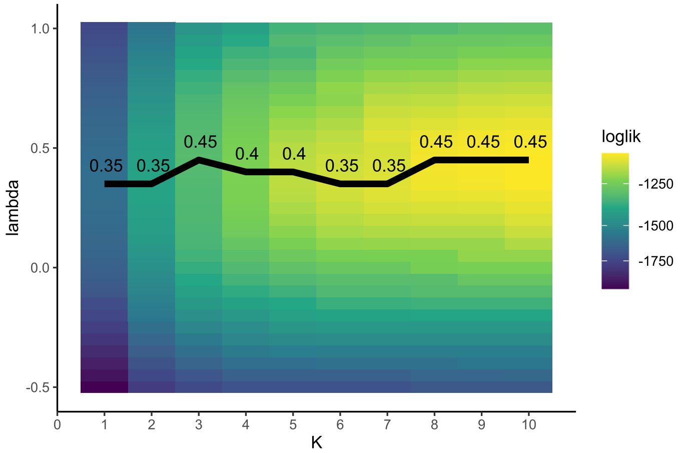

Fig. 3 Optimization using NPML approach. Log-likelihood values obtained from fitting Model (2) to the analysis data set with the NPML method, as implemented in the R package **boxcoxmix** [21]. The black line indicates the optimal values of the transformation parameter λ for varying numbers of mass points *K*. The corresponding values can be found in Table 2

also yielded an optimal AIC value at $\lambda_{opt} = 0.45$ (Fig. 2B).

The NPML approach led to similar results for the optimal value of λ in the setting without covariables. As seen in Fig. 3, the obtained values for the optimal λ ranged between 0.35 and 0.5. For a larger number of mass points ($K > 7$) the same optimal λ (= 0.45) as in the parametric approach was found.

Normality of the residuals Figure 4 shows the residual diagnostics obtained from fitting Model 6 to the analysis data, including hypercholesterolemia and hypertesnsion as risk factors. It is seen that even after transformation the fitted residuals were not normally distributed. However, homoscedasticity was better met after transformation with $\lambda_{opt} = 0.45$. Furthermore, the distribution of the residuals was less skewed after transformation.

Effects of risk factors As shown in Fig. 4, the residuals obtained from fitting Model 6 to the analysis data set did not perfectly follow a normal distribution, even after transformation of the response. Therefore, inference procedures that rely on asymptotic normality may not be the best choice to investigate the effects of risk factors on (transformed) GA size. To address this issue, we used a bootstrap approach to obtain the 95% confidence intervals of the parameters within Model (6). The results are presented in Table 3 and in Fig. 5. It is seen, that time was associated with the transformed GA size, growing by 0.42 (95% CI [0.36,0.50]) per year. Also the absence of hypercholesterolemia was associated with more rapid enlargement of the lesion size (estimate: 0.11, 95% CI [0.06,0.17]), while a slower progression in patients without hypertension (estimate: -0.09, 95% CI $[-0.17, -0.03]$) was found. Note that the estimated coefficients refer to transformed GA size and thus cannot be directly interpreted in terms of an enlargement of the GA size measured in mm^2.

Remark: Model fitting was performed on an imputed data set, using the R package **mice** [24] with one imputation. Results obtained from complete case analysis were almost identical.

Age-of-onset estimation

Figure 6 presents the estimated ages of disease onset of the study eyes, as obtained from Models (3) (without covariables) and (6) (with covariables). For the simple model

Table 2 Optimization using NPML approach

K	loglik	λ_{opt}	AIC$_{bm}$
1.00	-1593.02	0.35	3192.04
2.00	-1402.21	0.35	2814.41
3.00	-1310.29	0.45	2634.57
4.00	-1230.15	0.40	2478.30
5.00	-1164.85	0.40	2351.70
6.00	-1142.70	0.35	2311.39
7.00	-1127.57	0.35	2285.14
8.00	-1107.63	0.45	2249.26
9.00	-1102.10	0.45	2242.19
10.00	-1096.30	0.45	2234.60

The table presents the optimal values of the transformation parameter λ that were obtained from fitting Model (2) with the **boxcoxmix** package [21] using the analysis data set. In addition, the respective log-likelihood and AIC$_{bm}$ values (evaluated at the optimal λ values) are shown for varying numbers of mass points *K*. Following [21], the information criterion was defined as $\text{AIC}_{bm} = -2\log(L) + 2 \cdot (p + 2K)$. Hence the AIC values in the fourth column cannot be directly compared to the AIC values presented in Fig. 2

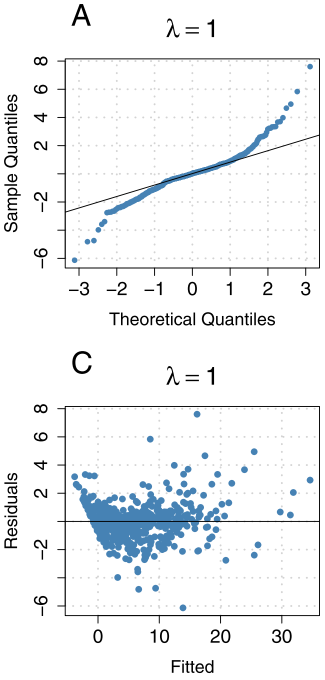

Fig. 4 Distribution of residuals. Residual diagnostics for Model (6) with transformation parameter $\lambda = 1$ (left column) and optimal transformation parameter $\lambda = 0.45$ (right column). Note that $\lambda = 1$ corresponds to a model with untransformed response. Panels **A** and **B** present normal quantile-quantile plots of the estimated residuals that were obtained from fitting Model (6) to the analysis data set. Panels **C** and **D** contain plots of estimated residuals vs. fitted values (fitted values include random effect terms)

without further covariables, the estimated mean age-of-onset was 66.93 (±7.56) years and for the model with covariables the estimated median age-of-onset was 67.21 (±6.49) years. This is in line with previously reported results, e.g. Li et al. [26] estimated the prevalence of GA in people under 64 years to range between 0.1% and 0.2%, depending on the country.

Estimation of GA size on the original scale

To obtain the distribution of GA size on the original scale, we sampled 10,000 times from the empirical distribution of the estimated residuals (obtained from Model (6)) and added these values to the fitted transformed GA size values $f_\lambda(y)$ before applying a reverse Box-Cox transformation. The back-transformed expected GA size values are shown in Fig. 7.

The root mean squared difference between the observed GA size and the modeled GA size was 1.10 mm^2, implying that estimated expected GA size values deviated by ca. 1 mm^2 on average from the true GA size values. The respective mean squared differences for alternative values of the transformation parameter λ are shown in Fig. 8.

Table 3 Analysis of risk factors in the analysis data set

Variable	Estimate	95% CI	*p*-value
time [in years]	0.42	(0.36, 0.50)	$<$ 0.0001
time x (hyperchol.= no)	0.11	(0.06, 0.17)	$<$ 0.0001
time x (hypertension = no)	-0.09	(-0.17 ,-0.03)	0.0004
Variance Term	**Estimate**		
Eye:Patient γ_i	1.83^2		
Patient ζ_j	4.03^2		
Residuals ϵ	0.42^2		

The table presents the coefficient estimates and bootstrap 95% confidence intervals that were obtained from fitting Model (6) with transformed response ($\lambda = 0.45$) to an imputed version of the analysis data set. The model parameter μ_Δ, which reflects the mean disease age at study entry, was estimated to be $\hat{\mu}_\Delta = 4.74$ (95% CI [3.41, 4.83]). *P*-values were obtained using the R package **lmerTest** [25]

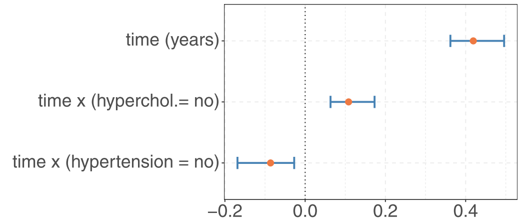

Fig. 5 Analysis of risk factors in the analysis data set. The plot shows the coefficient estimates (red dots) that were obtained from fitting Model (6) with transformed response ($\lambda = 0.45$) to an imputed version of the analysis data set. Bootstrap 95% confidence intervals are indicated by blue lines. For further details see Table 3

As can be seen here, the λ, that lead to a minimal difference on the original scale, was slightly larger than the optimal $\lambda = 0.45$ obtained by AIC-based methods. However, the variation in the average distances between observed and predicted values was rather small (minimal distance $1.05\,\mathrm{mm}^2$ at $\lambda = 0.55$, $1.06\,\mathrm{mm}^2$ at $\lambda = 0.50$, and $1.10\,\mathrm{mm}^2$ at $\lambda = 0.45$).

Prediction of next observation In clinical context, a prediction of the next observation of a patient already included in a clinical trial might be of interest. For each observed eye, for which values of more than three visits were present, we predicted the last observation. To this purpose we fitted a model to a training data set excluding the last observation while performance was measured on the last observation. The root mean squared difference between observed atrophy sizes and the mean predicted atrophy sizes was $\sqrt{\operatorname{avg}((\bar{\hat{y}} - y)^2)} = 1.67\,\mathrm{mm}^2$.

Discussion

Despite a high prevalence and extensive research efforts, there are currently no effective standard treatment options for GA. It is therefore essential to develop accurate models for disease progression that enable researchers to efficiently plan and design clinical trials.

In this article, we presented a comprehensive framework for modeling the course of GA size progression in longitudinal observational studies. Our modeling approach was derived from a linear enlargement model using transformed GA size as response variable. As shown in the Results section, the resulting model can be embedded in the class of linear mixed-effects models [27], allowing for the incorporation of risk factors, confounding variables, and measurements taken repeatedly from the same patients and eyes. Since the assumption of linear enlargement imposes numerous restrictions on the model parameters, it is necessary to adapt standard (unrestricted) mixed-effects modeling approaches to the specific structure of the proposed model. To this purpose, we developed

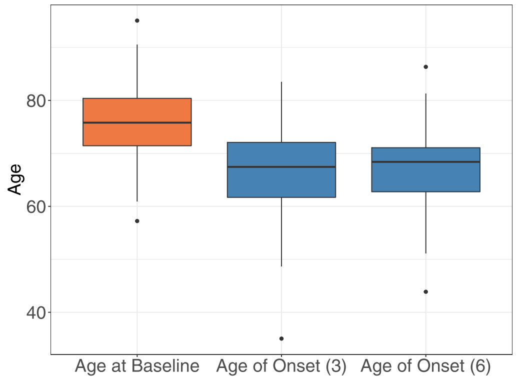

Fig. 6 Age-of-onset estimation. Boxplots of the observed ages at baseline (orange) and the modeled ages of disease onset (Models (3) and (6), blue) of the eyes in the analysis data set

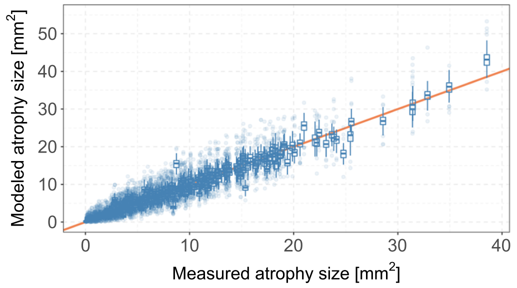

Fig. 7 Agreement between the estimated expected GA size values and the measured GA size values modeled distributions show one boxplot per observation and were generated by sampling from the residual distribution of Model (6), followed by a back-transformation of $z_{ijt} + \epsilon_u$ to the original scale. The orange line indicates a perfect fit

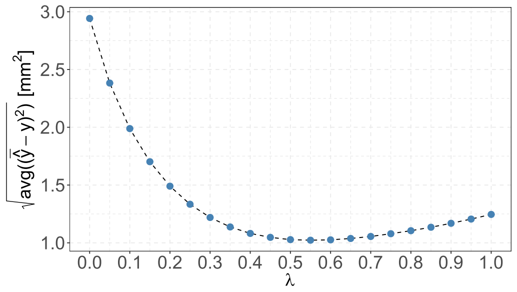

Fig. 8 Root mean squared deviation between modeled response and observed values for different transformation parameters. Deviation was measured by $\operatorname{avg}((\bar{\hat{y}} - y)^2)$. Here, the back-transformed estimate was defined by $\hat{y} = f^{-1}(\hat{y}^{(\lambda)} + \hat{\epsilon})$ (if f^{-1} existed), where f^{-1} was the inverse Box-Cox transformation, and $\bar{\hat{y}}$ was its mean (computed from sampling 10,000 times from the fitted residuals). For $\lambda = 0.45$, the deviation was $1.10\,\mathrm{mm}^2$ and a minimal squared deviation was reached at $\lambda = 0.55$ with a deviation of $1.05\,\mathrm{mm}^2$

an algorithm for GA size modeling that can be implemented using readily available software for fitting linear mixed-effects models.

To obtain the best transformation of GA size, we conducted a systematic search within the class of Box-Cox transformation models that included both parametric and non-parametric approaches. Our experiments yielded an optimal transformation that was close to the square-root function, thereby justifying earlier modeling strategies that assumed linear trajectories of square-root transformed GA size over time [18]. Of note, the square-root transformation has a straightforward interpretation in terms of a linear enlargement of the atrophy radius [15].

A convenient feature of the proposed modeling approach is that it yields estimates of the disease age of the eyes at study entry. This is important because patients can only be included in trials when the disease has already manifested. When applied to the analysis data set consisting of patients included in the FAM-study, disease age at study entry was estimated to range between 3.5 and 13.4 years (Model (6)). These estimates are in line with estimated prevalence values reported in the literature [4], but the resulting ages of disease onset were smaller than previously modeled ages using data partly from the same study [28].

Since the proposed modeling approach employs a transformed response variable, care has to be taken when making predictions of future values of atrophy size. As argued in the Results section, predictions with a naive back-transformation may show a bias due to the nonlinearity of the square-root function. To address this issue, we proposed a sampling approach that allows for drawing valid conclusions and making undistorted predictions of GA size on its original scale. In the analysis data set, estimated expected GA size values derived from the proposed model deviated 1.10 mm^2 on average from the respective observed values.

Generally, the model proposed here allows for performing statistical hypothesis tests on a set of risk factors suspected to accelerate or slow down GA size enlargement. This strategy was illustrated in the Results section, where an analysis of a GA patient sample of the FAM study identified significant interaction effects between hypercholesterolemia, hypertension and time. Although a number of studies have shown a link between cardiovascular risk factors and AMD, the role of hypertension, atherosclerosis, high BMI, diabetes mellitus, higher plasma fibrinogen and hyperlipidaemia remain equivocal owing to inconsistent findings (reviewed in [29]). High blood pressure is shown to be associated with lower choroidal blood flow and disturbed vascular homeostasis [30]. Since perfusion deficits in the choriocapillaris, the innermost layer of the choroid, are associated with future GA progression [31], an associate between hypertension and increased GA progression appears biologically plausible. Regarding the association of hypercholesterinemia and decreased GA progression, the biological plausibility remains elusive. The majority of previous studies did not find any relationship between systemic cholesterol levels and progression to early AMD, GA or nAMD (reviewed in [29]), although two studies found an association between serum cholesterol on the development of late stage AMD [32, 33]. Interestingly, one of these studies reported that serum cholesterol levels have a protective effect on the development of nAMD, while they are a risk factor for the development of GA [32]. These observations apparently are in contrast to our results; however, there is evidence that different mechanisms may be involved in driving GA enlargement than those increasing the risk of de novo GA development [6]. Further validation of the risk factors, especially on an external data set, is necessary

While it has been established that so-called nascent GA progresses to manifest GA [34], the trajectory of early GA – prior to the minimum lesion size requirement for clinical trials (e.g., 2.5 mm^2) – is poorly understood. The information derived by this modeling strategy can be used to design future intervention studies, for example regarding the stratification of patient groups and the definition of inclusion criteria. Of note, the proposed modeling approach is not restricted to established epidemiological covariables like hypertension but may also incorporate novel markers of disease progression such as patient-reported outcome measures [35], digital biomarkers, and machine-learning-based scores derived from structural imaging data [36]. The proposed model constitutes a flexible framework to systematically investigate the transition from intermediate to late AMD in large observational studies such as the MACUSTAR study (ClinicalTrials.gov: NCT03349801) [37].

Conclusions

We have provided a comprehensive framework for modelling the trajectories of uni- or bilateral Ga size progression in longitudinal observational studies. Our analysis shows that a square-root transformation of atropy size is recommended before model fitting. The proposed modelling approach allows for the estimation of age-of-onset, identification of risk factors and prediction of future GA size. The risk factors analyzed here require further validation in an external study population.

Abbreviations

AIC: Akaike's information criterion; AMD: Age related macular degeneration; avg: Average; BMI: Body max index; CI: Confidence interval; DSGA: Directional spread in geographic atrophy study; FAF: Fundus autoflourescence; FAM: Fundus autoflourescence in AMD study; GA: Geographic atrophy; IQR: Interquartile range; ML: Maximum likelihood; nAMD: Neovascular AMD; NPML: Nonparametic maximum likelihood; RPE: Retinal pigment epithelium

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-021-01356-0.

Additional file 1: Supplementary information.

Acknowledgements

This work was supported by the DFG (German Research Foundation) Grant No FL658/4-1 and FL658/4-2 (MF); PF950/1-1 (MP) Research Priority Program Age-Related Macular Degeneration Grant SPP 1088, Ho 1926/1-3 (FGH); Grants from National Institutes of Health Core Grant (EY014800) (MF); and in part by an Unrestricted Grant from Research to Prevent Blindness, New York, NY, to the Department of Ophthalmology & Visual Sciences, University of Utah. The authors would like to thank Joanna Czauderna for her support in conducting the DSGA study.

Authors' contributions

Conception and design: MS, MF, CB. Implementation and analysis: CB. Data collection and interpretation: MF, MP, LG, ML, SSV. Review and approval of final manuscript: CA, MP, MF. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

The datasets generated during and/or analysed during the current study are not publicly available in order to protect the privacy of study participants. However, they are available from the principal investigators of the FAM and DSGA studies on reasonable scientific request.

Declarations

Ethics approval and consent to participate

This study does not include human subjects. Data presented here are from the *Fundus Autofluorescence in AMD* (FAM) study (ClinicalTrials.gov Identifier: NCT00393692) and the *Directional Spread in Geographic Atrophy* (DSGA) study (ClinicalTrials.gov Identifier: NCT02051998). The FAM study was approved by the human ethics committees at the University of Bonn and the local Institutional Review Boards and the local ethics committees at the study centers (University of Bonn, University of Heidelberg, University of Leipzig, Ludwig-Maximilians-University Munich, St. Franziskus Hospital Münster, University of Würzburg). The human ethics committees at the University of Bonn approved the DSGA study. All research adhered to the tenets of the Declaration of Helsinki. All participants provided informed consent.

Consent for publication

All authors have read and approved the submission of the manuscript.

Competing interests

Financial disclosures:

S. Schmitz-Valckenberg reports grants from Acucela/Kubota Vision, personal fees from Apellis, grants and personal fees from Novartis, grants and personal fees from Allergan, grants and personal fees from Bayer, grants and personal fees from Bioeq/Formycon, grants, personal fees and non-financial support from Carl Zeiss MediTec AG, grants and non-financial support from Centervue, personal fees from Galimedix, grants, personal fees and non-financial support from Heidelberg Engineering, grants from Katairo, non-financial support from Optos, personal fees from Oxurion, outside the submitted work.

M. Fleckenstein reports grants, personal fees and non-financial support from Heidelberg Engineering, non-financial support from Zeiss Meditech, grants and non-financial support from Optos, personal fees and grant from Novartis, personal fees from Bayer, grants and personal fees from Genentech, from Roche, outside the submitted work; In addition, Dr. Fleckenstein has a patent US20140303013 A1 pending.

F.G. Holz reports grants and personal fees from Acucela, Allergan, Apellis, Bayer, Bioeq/Formycon, Roche/Genentech, Geuder, Heidelberg Engineering, ivericbio, Kanghong, Novartis, Zeiss; personal fees from Boehringer-Ingelheim, Grayburg Vision, LinBioscience, Pixium Vision, Stealth BioTherapeutics, Aerie, Oxurion outside the submitted work.

M. Schmid reports personal fees from Pixium Vision.

The remaining authors have no competing interest.

Author details

[1]Department of Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany. [2]John A. Moran Eye Center, University of Utah, Salt Lake City, USA. [3]Ophthalmic Genetics and Visual Function Branch, National Eye Institute, Bethesda, MD, USA. [4]Institute for Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-University, Munich, Germany. [5]Department of Ophthalmology, University Hospital Bonn, Bonn, Germany.

Received: 29 December 2020 Accepted: 22 July 2021
Published online: 17 August 2021

References

1. Lim LS, Mitchell P, Seddon JM, Holz FG, Wong TY. Age-related macular degeneration. Lancet. 20121728–38.
2. Wong WL, Su X, Li X, Cheung CMG, Klein R, Cheng C-Y, Wong TY. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. Lancet Glob Health. 2014;2:106–16.
3. Sunness JS. The natural history of geographic atrophy, the advanced atrophic form of age-related macular degeneration. Mol Vis. 1999;5:25.
4. Friedman DS, O'Colmain BJ, Munoz B, Tomany SC, McCarty C, De Jong P, Nemesure B, Mitchell P, Kempen J, et al. Prevalence of age-related macular degeneration in the united states. Arch Ophthalmol. 2004;122: 564–72.
5. Holz FG, Schmitz-Valckenberg S, Fleckenstein M. Recent developments in the treatment of age-related macular degeneration. J Clin Investig. 2014;124:1430–8.
6. Fleckenstein M, Mitchell P, Freund KB, Sadda S, Holz FG, Brittain C, Henry EC, Ferrara D. The progression of geographic atrophy secondary to age-related macular degeneration. Ophthalmology. 2018;125:369–90.
7. Sunness JS, Gonzalez-Baron J, Applegate CA, Bressler NM, Tian Y, Hawkins B, Barron Y, Bergman A. Enlargement of atrophy and visual acuity loss in the geographic atrophy form of age-related macular degeneration. Ophthalmology. 1999;106:1768–79.
8. Holz FG, Sadda SR, Staurenghi G, Lindner M, Bird AC, Blodi BA, Bottoni F, Chakravarthy U, Chew EY, Csaky K, et al. Imaging protocols in clinical studies in advanced age-related macular degeneration: recommendations from classification of atrophy consensus meetings. Ophthalmology. 2017;124:464–78.
9. Arslan J, Samarasinghe G, Benke KK, Sowmya A, Wu Z, Guymer RH, Baird PN. Artificial intelligence algorithms for analysis of geographic atrophy: A review and evaluation. Transl Vis Sci Technol. 2020;9(2):57.
10. Holekamp N, Wykoff CC, Schmitz-Valckenberg S, Monés J, Souied EH, Lin H, Rabena MD, Cantrell RA, Henry EC, Tang F, et al. Natural history of geographic atrophy secondary to age-related macular degeneration: results from the prospective proxima a and b clinical trials. Ophthalmology. 2020;127(6):769–83.
11. Keenan TD, Agrón E, Domalpally A, Clemons TE, van Asten F, Wong WT, Danis RG, Sadda S, Rosenfeld PJ, Klein ML, et al. Progression of geographic atrophy in age-related macular degeneration: Areds2 report number 16. Ophthalmology. 2018;125:1913–28.
12. Schmitz-Valckenberg S, Sahel J-A, Danis R, Fleckenstein M, Jaffe GJ, Wolf S, Pruente C, Holz FG. Natural history of geographic atrophy progression secondary to age-related macular degeneration (geographic atrophy progression study). Ophthalmology. 2016;123:361–8.
13. Dreyhaupt J, Mansmann U, Pritsch M, Dolar-Szczasny J, Bindewald A, Holz F. Modelling the natural history of geographic atrophy in patients with age-related macular degeneration. Ophthalmic Epidemiol. 2005;12: 353–62.
14. Holz FG, Bindewald-Wittich A, Fleckenstein M, Dreyhaupt J, Scholl HP, Schmitz-Valckenberg S, Group F-S, et al. Progression of geographic atrophy and impact of fundus autofluorescence patterns in age-related macular degeneration. Am J Ophthalmology. 2007;143:463–72.
15. Feuer WJ, Yehoshua Z, Gregori G, Penha FM, Chew EY, Ferris FL, Clemons TE, Lindblad AS, Rosenfeld PJ. Square root transformation of geographic atrophy area measurements to eliminate dependence of growth rates on baseline lesion measurements: a reanalysis of age-related eye disease study report no. 26. JAMA Ophthalmology. 2013;131:110–1.
16. Box GE, Cox DR. An analysis of transformations. J R Stat Soc Ser B (Methodol). 1964;26:211–43.

17. Lindner M, Bezatis A, Czauderna J, Becker E, Brinkmann CK, Schmitz-Valckenberg S, Fimmers R, Holz FG, Fleckenstein M. Choroidal thickness in geographic atrophy secondary to age-related macular degeneration. Invest Ophthalmol Vis Sci. 2015;56:875–82.
18. Shen L, Liu F, Nardini HG, Del Priore LV. Natural history of geographic atrophy in untreated eyes with nonexudative age-related macular degeneration: a systematic review and meta-analysis. Ophthalmol Retin. 2018;2:914–21.
19. Akaike H. A new look at the statistical model identification. IEEE Trans Autom Control. 1974;19(6):716–23.
20. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;67:1–48. https://doi.org/10.18637/jss.v067.i01.
21. Almohaimeed A, Einbeck J, Almohaimeed MA. Package Boxcoxmix. 2018. R package version 0.21. https://CRAN.R-project.org/package=boxcoxmix. Last accessed: 08 Feb 2021.
22. Sadda SR, Guymer R, Holz FG, Schmitz-Valckenberg S, Curcio CA, Bird AC, Blodi BA, Bottoni F, Chakravarthy U, Chew EY, et al. Consensus definition for atrophy associated with age-related macular degeneration on oct: classification of atrophy report 3. Ophthalmology. 2018;125(4):537–48.
23. Godunova EK. Jensen inequality. Encyclopedia of Mathematics: Springer Verlag GmbH, European Mathematical Society; 2011. http://encyclopediaofmath.org/index.php?title=Jensen_inequality&oldid=47465. Accessed 08 Feb 2020.
24. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in r. J Stat Softw. 2011;45:1–67.
25. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: Tests in linear mixed effects models. J Stat Softw. 2017;82(13):1–26. https://doi.org/10.18637/jss.v082.i13.
26. Li JQ, Welchowski T, Schmid M, Mauschitz MM, Holz FG, Finger RP. Prevalence and incidence of age-related macular degeneration in europe: a systematic review and meta-analysis. Br J Ophthalmol. 2020;104(8):1077–84.
27. Verbeke G, Molenberghs G. Linear Mixed Models for Longitudinal Data. Berlin Heidelberg: Springer Science & Business Media; 2009.
28. Adrion C, Fleckenstein M, Schmitz-Valckenberg S, Holz F, Mansmann U. Estimation of disease onset for patients with geographic atrophy due to age-related macular degeneration. Gesundheitswesen. 2010;72:207.
29. Heesterbeek TJ, Lorés-Motta L, Hoyng CB, Lechanteur YT, den Hollander AI. Risk factors for progression of age-related macular degeneration. Ophthalmic Physiol Opt. 2020;40(2):140–70.
30. Metelitsina T, Grunwald J, DuPont J, Ying G. Effect of systemic hypertension on foveolar choroidal blood flow in age related macular degeneration. Br J Ophthalmol. 2006;90(3):342–6.
31. Müller PL, Pfau M, Schmitz-Valckenberg S, Fleckenstein M, Holz FG. Optical Coherence Tomography-Angiography in Geographic Atrophy. Ophthalmologica. 2020;244(1):42–50. https://doi.org/10.1159/000510727.
32. Tomany S, Wang JJ, Van Leeuwen R, Klein R, Mitchell P, Vingerling JR, Klein BE, Smith W, De Jong PT. Risk factors for incident age-related macular degeneration: pooled findings from 3 continents. Ophthalmology. 2004;111(7):1280–7. https://doi.org/10.1016/j.ophtha.2003.11.010.
33. Buch H, Vinding T, La Cour M, Jensen GB, Prause JU, Nielsen NV. Risk factors for age-related maculopathy in a 14-year follow-up study: the copenhagen city eye study. Acta Ophthalmol Scand. 2005;83(4):409–18.
34. Wu Z, Luu CD, Hodgson LA, Caruso E, Tindill N, Aung KZ, McGuinness MB, Makeyeva G, Chen FK, Chakravarthy U, et al. Prospective longitudinal evaluation of nascent geographic atrophy in age-related macular degeneration. Ophthalmol Retin. 2020;4(6):568–75.
35. Ying G-S, Maguire MG. Complications of Age-related Macular Degeneration Prevention Trial Research Group. Development of a risk score for geographic atrophy in complications of the age-related macular degeneration prevention trial. Ophthalmology. 2011;118(2):332–8. https://doi.org/10.1016/j.ophtha.2010.06.030.
36. Pfau M, von der Emde L, de Sisternes L, Hallak JA, Leng T, Schmitz-Valckenberg S, Holz FG, Fleckenstein M, Rubin DL. Progression of photoreceptor degeneration in geographic atrophy secondary to age-related macular degeneration. JAMA Ophthalmol. 2020;138(10):1026–34.
37. Finger RP, Schmitz-Valckenberg S, Schmid M, Rubin GS, Dunbar H, Tufail A, Crabb DP, Binns A, Sánchez CI, Margaron P, et al. Macustar: development and clinical validation of functional, structural, and patient-reported endpoints in intermediate age-related macular degeneration. Ophthalmologica. 2019;241(2):61–72.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

3.3 Publication C: An imputation approach using subdistribution weights for deep survival analysis with competing events

Gorgi Zadeh S, Behning C, Schmid M. An imputation approach using subdistribution weights for deep survival analysis with competing events. In: Scientific Reports 2022; 12 (1): 3815 https://doi.org/10.1038/s41598-022-07828-7

Implementations are available on GitHub:
https://github.com/shekoufeh/Deep-Survival-Analysis-With-Competing-Events

scientific reports

OPEN

An imputation approach using subdistribution weights for deep survival analysis with competing events

Shekoufeh Gorgi Zadeh✉, Charlotte Behning & Matthias Schmid

With the popularity of deep neural networks (DNNs) in recent years, many researchers have proposed DNNs for the analysis of survival data (time-to-event data). These networks learn the distribution of survival times directly from the predictor variables without making strong assumptions on the underlying stochastic process. In survival analysis, it is common to observe several types of events, also called competing events. The occurrences of these competing events are usually not independent of one another and have to be incorporated in the modeling process in addition to censoring. In classical survival analysis, a popular method to incorporate competing events is the subdistribution hazard model, which is usually fitted using weighted Cox regression. In the DNN framework, only few architectures have been proposed to model the distribution of time to a specific event in a competing events situation. These architectures are characterized by a separate subnetwork/pathway per event, leading to large networks with huge amounts of parameters that may become difficult to train. In this work, we propose a novel imputation strategy for data preprocessing that incorporates weights derived from a time-discrete version of the classical subdistribution hazard model. With this, it is no longer necessary to add multiple subnetworks to the DNN to handle competing events. Our experiments on synthetic and real-world datasets show that DNNs with multiple subnetworks per event can simply be replaced by a DNN designed for a single-event analysis without loss in accuracy.

In the recent years deep networks have become the state-of-the-art method in various applications, for instance in object detection [1], image captioning [2], image classification[3,4], speech recognition [5], and many other areas. One key advantage of deep neural networks is their capacity to learn specific intermediate representations/features of the data in a hierarchical manner[6] in order to create a mapping from the input predictor variables onto the outcome. In addition to other novel machine learning methods developed for survival analysis[7], recently, there has been a growing interest in using deep neural networks for this purpose, see for example, the works by Giunchiglia et al.[8], Lee et al.[9], Zafar Nezhad et al.[10] and many others[11–16].

In survival analysis the outcome is usually defined by the time duration until one or more events occur[17]. For instance in the medical field this event could be recurrence of a disease or patient's death after an intervention. A multitude of examples can e.g. be found in the work by Lee et al.[18]. Since survival data (also called *time-to-event* data) are collected over time, they are often subject to right censoring, which means that the event times of some instances are only known up to a minimum survival time. The real event times of these instances remain unknown as they are no longer observed beyond the time of censoring. Often, right censoring occurs when patients drop out of a study or when patients have not experienced any event before study end.

Many observational studies track more than one event. Often these so-called *competing events* do not occur independently, and therefore require to be analyzed together in order to avoid bias. For instance, in the CRASH-2 trial[19], which is a large randomized study on hospital death in adult trauma patients, there are multiple recorded causes of death throughout the study. The causes include death due to bleeding, head injury, multi-organ failure and others. Obviously, the occurrences of these causes are not independent. More examples on competing risks data can be found in the works by Lau et al.[20] and Austin et al.[21].

For modeling the time span until a specific event of type $j \in \{1, \ldots, J\}$ occurs, multiple approaches have been proposed. For example, Prentice et al.[22] model the *cause-specific hazard functions* of each event separately as $\xi_j(t|x) = \lim_{\Delta t \longrightarrow 0}\{P(t \leq T < t + \Delta t, \epsilon = j \mid T \geq t, x)/\Delta t\}$, where $x = (x_1, \ldots, x_p)^T$ is the vector of

Department of Medical Biometry, Informatics and Epidemiology, Faculty of Medicine, University of Bonn, Sigmund-Freud-Str: 25, D-53127 Bonn, Germany. ✉email: shekoufeh.gorgizadeh@imbie.uni-bonn.de

time-constant predictor variables and ϵ is a random variable indicating the type of the event that occurs at the first observed event time T. In their approach, a separate model is used for each ξ_j, treating the individuals that experience any of the respective competing events as censored. Another approach, on which the methods considered in this paper are based, is the subdistribution hazard model by Fine and Gray[23]. This approach aims at modeling the *cumulative incidence functions* $F_j(t|x) = P(T \leq t, \epsilon = j \,|\, x)$. For any event j of interest, the model considers a *subdistribution hazard* function $\lambda_j(t|x) = \lim_{\Delta t \longrightarrow 0}\{P(t \leq \vartheta_j < t + \Delta t \,|\, \vartheta_j \geq t, x)/\Delta t\}$, where ϑ_j is the "subdistribution time" defined by $\vartheta_j = T$ if $\epsilon = j$ and $\vartheta_j = \infty$ otherwise. Thus, ϑ_j corresponds to the time to the occurrence of a type-j event, assuming that such an event can never be observed (i.e. $\vartheta_j = \infty$) if a competing event occurs first. It can be shown[23] that specifying a regression model for $\lambda_j(t|x)$ allows for modeling cumulative incidences of type-j events without having to model the hazard functions of the other events. Thus, only one subdistribution hazard model is required if the interest is in the cumulative incidence function of the type-j event. This is unlike cause-specific hazards modeling, where all $\xi_1, \ldots, \xi_J$ need to be considered together to calculate cumulative incidence probabilities.

To analyze competing events data using deep neural networks, Lee et al.[9] proposed the DeepHit network that directly learns the distribution of survival times for an event of interest while at the same time handling the competing events. In their architecture, a separate subnetwork is added for each competing event. Similarly, Gupta et al.[11] use separate subnetworks per event. In another work, Nagpal et al.[24] proposed a Deep Survival Machine (DSM) to learn a mixture of parametric distributions (e.g. Weibull or log-normal) for estimating the conditional survival function $S(t|x) = P(T > t|x)$. Again, in this model an additional set of parameters is added to describe the event distribution for each competing risk.

In this work, instead of extending a network's architecture by multiple subnetworks to handle competing events, we follow the approach by Fine and Gray and propose to employ deep network architectures for a *single* event of interest[8,25–27]. To incorporate competing events, our method works on input data that have been preprocessed using an imputation strategy based on subdistribution weights (see Methods section for details). As will be demonstrated, this strategy allows analysts to benefit from the advantages of existing single-event implementations for time-to-event data (particularly, from much simpler architectures with smaller sets of parameters) while being able to avoid a possible bias caused by ignoring competing events. In our experiments on simulated and real-world datasets, we show that approximately the same performance can be gained without the need for specifying a complex network architecture with multiple event-specific parameter sets.

The key contributions of this work are: (1) We propose a novel preprocessing strategy for deep survival networks that enables a valid analysis of competing-risks data, even if the respective network architecture was originally designed to handle one event only. (2) We demonstrate the feasibility of our approach by comparing two variants of the DeepHit architecture. Specifically, we compare a DeepHit model with *two* subnetworks (designed to analyze the original input data with two competing events) to a DeepHit model with only *one* subnetwork (designed to analyze one event of interest and based on a modified input data set that was preprocessed using our imputation method). (3) Using simulations, we analyze the behavior of deep survival architectures that are designed to analyze one event of interest. Specifically, we demonstrate that these architectures perform better (in terms of both calibration and discrimination) when the proposed preprocessing strategy is applied than when the original input data (treating observations with a competing event as censored) are used.

Methods

Notations and definitions. To be able to use single-event DNN architectures like DeepSurv[25], SurvivalNet[26], RNN-Surv[8] and DRSA[27], continuous survival and censoring times have to be grouped. To this end, we define time intervals $[0, a_1), [a_1, a_2), ..., [a_{k-1}, \infty)$, where k is a natural number. Further denote by $T_i \in \{1, ..., k\}$ and $C_i \in \{1, ..., k\}$ the resulting discrete event and censoring times, respectively, of an individual contained in an i.i.d. sample of size $n, i = 1, \ldots, n$. In this definition, $T_i = t$ means that the event has happened in time interval $[a_{t-1}, a_t)$. It is assumed that T_i and C_i are independent random variables ("random censoring"). Furthermore, it is assumed that the censoring time does not depend on the parameters used to model the event time, i.e. the censoring mechanism is "non-informative" for T_i[22,28]. For right-censored data, the observed time is defined by $\tilde{T}_i = \min(T_i, C_i)$, i.e. $\tilde{T}_i$ corresponds to true event time if $T_i \leq C_i$, and to the censoring time otherwise. The random variable $\Delta_i := I(T_i \leq C_i)$ indicates whether $\tilde{T}_i$ is right-censored ($\Delta_i = 0$) or not ($\Delta_i = 1$). In addition to the event of interest (defined without loss of generality by $j = 1$), we assume that each individual can experience one out of $J - 1$ competing events, $j \in \{2, \ldots, J\}$. The type of event that the i-th individual experiences at T_i is represented by the random variable $\epsilon_i \in \{1, ..., J\}$[29]. The values of the predictor variables of the i-th individual are denoted by $x_i = (x_{i1}, \ldots, x_{ip})^T$. Analogous to the works by Fine and Gray[23] and Berger et al.[30], we are interested in modeling the cumulative incidence function $F_1(t|x) = P(T \leq t, \epsilon = 1 \,|\, x)$ of a type-1 event using the subdistribution hazard approach described above. To fit their proposed models, both Fine & Gray and Berger et al. considered the optimization of *weighted* versions of the underlying partial and binomial log-likelihood functions. While these techniques turn out to be highly effective when fitting parametric models to sets of lower-dimensional data, it is challenging to adapt them to learning tasks involving deep survival models. Specifically, the method by Fine & Gray relies on a continuous time scale and does not apply directly to the discrete (grouped) event times specified above. On the other hand, the method by Berger et al., which extends the Fine & Gray method to discrete event times, requires the input data to be "augmented" to up to $n \cdot k$ instances, implying a potentially huge increase in dimension. Clearly, this approach is not feasible for deep learning tasks, which typically rely on large values of n. We propose to address the aforementioned challenges by specifying a preprocessing strategy that operates directly on the discrete event times, while at the same time preserving the dimension of the input data.

Imputation strategy. In this section we describe the imputation strategy to preprocess the time-discrete input data. The aim is to modify the data such that it is possible to obtain valid estimates of the cumulative incidence function $F_1(t|x)$ by training a single-event DNN. As outlined in the Introduction section, training could be based on the specification of a subdistribution time $\vartheta \equiv \vartheta_1$, which could be subsequently used to learn a single-event DNN with input data $(\min(\vartheta_i, C_i), I(\vartheta_i \leq C_i), x_i^\top)$, $i = 1, \ldots, n$. A problem of this strategy is that it cannot be readily applied in practice, as the aforementioned input data are partly unknown. We therefore propose to apply additional preprocessing steps to the available input data. The details are as follows:

First, consider those individuals i with $\Delta_i \epsilon_i \in \{0, 1\}$. Clearly, it is not necessary to preprocess the input data of these individuals, since both $\min(\vartheta_i, C_i) = \tilde{T}_i$ and $I(\vartheta_i \leq C_i) = \Delta_i$ are known in these cases. Next, consider those individuals who experience a competing event first, i.e. $\Delta_i \epsilon_i > 1$. For these individuals $\vartheta_i = \infty$, so that $I(\vartheta_i \leq C_i) = 0$ is known. However, $\min(\vartheta_i, C_i) = \min(\infty, C_i) = C_i$ is unknown in these cases due to the fact that the value of the censoring time C_i is unobserved.

The main idea of our approach is, therefore, to impute the missing values of C_i by sampling a censoring time for any individual i who experiences a competing event first. Our strategy is as follows:

(i) Following Berger et al.[30], we first define the set of discrete *subdistribution weights*$u_{it} = I(t \leq \min(\vartheta_i, C_i))$, $i = 1, \ldots, n, t = 1, \ldots, k - 1$, indicating whether individual i is at risk of a type-1 event at time point t ($u_{it} = 1$) or not ($u_{it} = 0$). We further denote by $r(t)$ the *risk set* of individuals who have neither experienced a type-1 event nor have been censored before t. As outlined above, $r(t)$ is not fully known for individuals who experience a competing event first. These individuals remain at risk beyond $\tilde{T}_i$ until eventually they experience the censoring event.

(ii) In line with Fine & Gray[23] and Berger et al.[30], we specify an *estimate* of the subdistribution weights that can be computed from the available data. Denoting this estimate by w_{it}, $i = 1, \ldots, n$, $t = 1, \ldots, k - 1$, we set $w_{it} = 1$ if $t \leq \tilde{T}_i$, knowing that individuals remain at risk (i.e. belong to $r(t)$) until $\tilde{T}_i$. For $t > \tilde{T}_i$ and $\Delta_i \epsilon_i > 1$, we estimate u_{it} by the conditional probability of individual i being part of $r(t)$, given knowledge that it is part of $r(\tilde{T}_i)$. This conditional probability can in turn be estimated by

$$w_{it} := \frac{\hat{G}(t-1)}{\hat{G}(\tilde{T}_i - 1)}, \quad \tilde{T}_i < t \leq k - 1, \tag{1}$$

where $\hat{G}(t)$ is an estimate of the censoring survival function $G(t) = P(C_i > t)$. For the experiments in this paper, we used the R package *discSurv*[31], which implements a nonparametric life table estimator to obtain estimates of $G(t)$.

(iii) In the final step, we use w_{it} to sample estimates of the censoring times of individuals who experience a competing event first. For this, we generate random numbers $\hat{C}_i$ from discrete distributions with supports $(\tilde{T}_i + 1, \ldots, k - 1)$ that are defined by $P(\hat{C}_i = t) = \Delta w_{it}$, where $\Delta w_{it} = w_{it-1} - w_{it}$. The so-obtained numbers are subsequently used to impute the unobserved values $\min(\vartheta_i, C_i)$. A visualization of the proposed imputation strategy is presented in Fig. 1.

Note that our method bears some similarities to the work by Ruan and Gray[32], who suggested a multiple imputation approach to model continuous-time survival data in a non-DNN context. The preprocessing strategy proposed here differs from Ruan and Gray[32] in three aspects: First, Ruan and Gray considered models in continuous time, whereas the DNN architectures considered here operate on a discrete time scale. Accordingly, Ruan and Gray used a conditional Kaplan-Meier estimator to estimate the censoring distribution, implying that the resulting weight differences Δw_{it} occur at random time points (whereas we consider fixed [user-specified] interval borders $a_1 < a_2 < \ldots < a_{k-1}$ to define Δw_{it}). Second, Ruan and Gray proposed to estimate their quantities of interest (e.g. the parameters of a proportional subdistribution hazard model and/or cumulative incidence functions at selected time points) by applying a multiple imputation strategy. Accordingly, the authors proposed to generate multiple imputed data sets and to average estimates from the respective (multiple) analyses based on the imputed data. This is in contrast to our approach, which assumes that DNN architectures are able to capture the relevant aspects of the data-generating process using a single imputation only. Third, Ruan and Gray mostly focus on semiparametric survival models in a non-machine-learning context ("allowing standard software to be used for the analysis"), whereas the focus of this work is on the nonparametric estimation of cumulative incidence functions using DNN architectures with potentially higher-dimensional predictor spaces.

In the next section we demonstrate that without loss of accuracy, the use of the imputed data simplifies the analysis of competing-risks data by training single-event DNNs.

Experimental analysis

DeepHit network. To investigate the effectiveness of the proposed method, we used the DeepHit architecture by Lee et al.[9]. DeepHit is a DNN that allows to have a learnable survival function that maps the predictor variables vector x_i into a probability distribution vector $\mathbf{y}_i = \left[y_{1,1}, \ldots, y_{1,k}, \ldots, y_{J,1}, \ldots, y_{J,k}\right]$. In this vector, element $y_{\epsilon,t}$ is the estimated probability that instance i with predictor variables x_i will experience the ϵth event at time point t. Through non-linear activation functions, DNNs, and in particular DeepHit can learn potentially non-linear, even non-proportional, relationships between the predictor variables and the events[9]. A fully connected layer consists of neurons connected to all neurons in the adjacent layer. Each neuron works as a simple linear classifier ($h = f\left(\sum v_m x_m\right)$, where h is the output, v_m is the network weight, x_m the input from the mth neuron in the previous layer, and f is the activation function) that receives input from the neurons in the previous layer and sends output to every neuron in the next layer. DeepHit consists of a "shared sub-network" that has two fully connected layers. (Note that in the work by Lee et al.[9], the authors use one fully connected layer for their experiments. However, empirically we found that using two fully connected layers improves the overall accuracy.) The shared sub-network creates an intermediate representation that is further combined with the

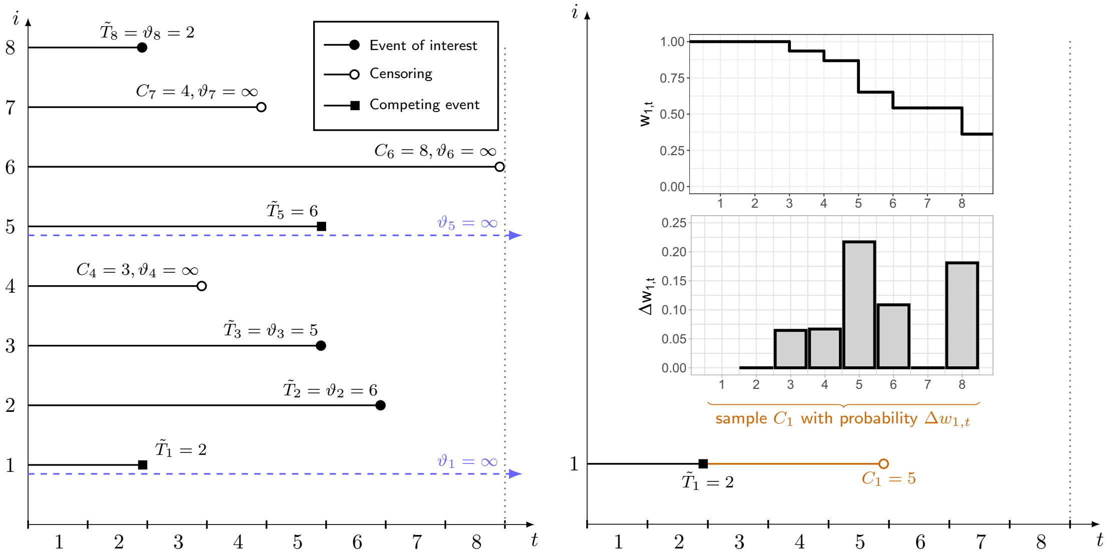

Figure 1. Illustration of the imputation strategy. The left panel presents the subdistribution times of eight randomly sampled individuals. Individuals 1 and 5 experienced the competing event first, implying that their censoring times are unobserved (as illustrated by the time span for $i = 1$ in the right panel). For these individuals, censoring times are estimated by first calculating estimated subdistribution hazard weights $w_{i,t}$ (see upper right diagram). From that, the weight differences $\Delta w_{i,t}$ are calculated and used to sample censoring times C_i, which are in turn used to impute the unobserved values of $C_i = \min(C_i, \vartheta_i)$. Note that the bars in the lower right panel correspond to the heights of the steps in the upper right panel.

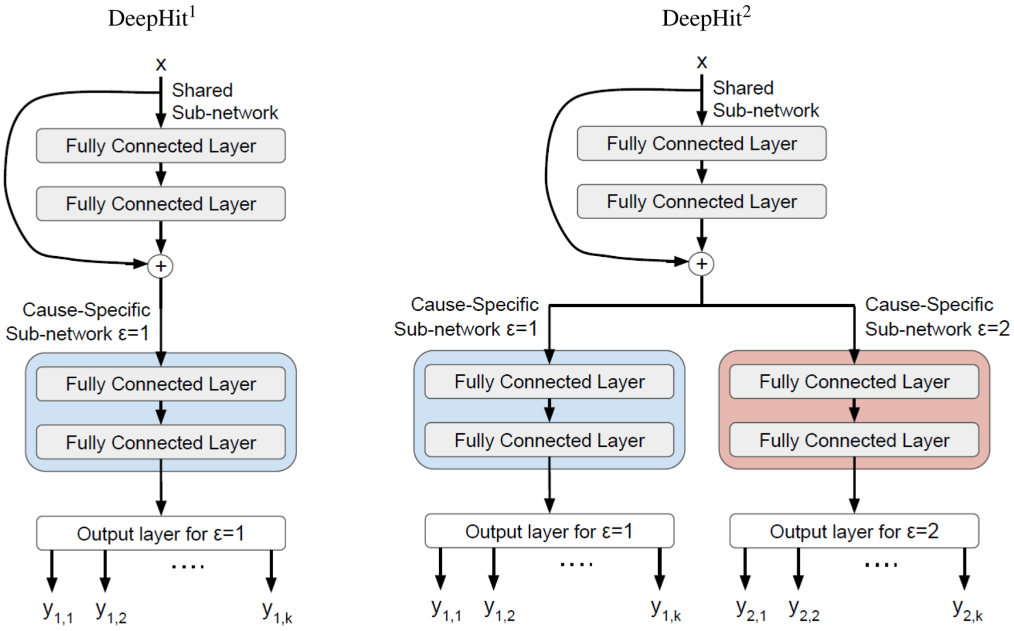

Figure 2. Visualization of the DeepHit[1] and DeepHit[2] architectures used in the experiments.

input features and passed on to J "cause-specific sub-networks". As recommended by Lee et al.[9], we used two fully connected layers in each sub-network. The output of each cause-specific sub-network is a vector that estimates the probability of the first hitting time of a specific cause j at each time point t (see Fig. 2). For training DeepHit, the authors use the log-likelihood of the joint distribution of the first hitting time as well as another loss term to incorporate a mixture of cause-specific ranking loss functions. They also modified the loss to handle right-censored data. In our experiments, we use the same loss term that was used to optimize DeepHit[9].

To assess the performance of our proposed method we compared three different setups: (1) *New approach using single-event DNN with preprocessed input data:* We trained the DeepHit network with only one subnetwork (see Fig. 2, DeepHit^1). Instead of the original input data, we used the modified version of the input data (with T_i replaced by ϑ_i), in which the censoring times corresponding to individuals with observed competing events were imputed using the subdistribution weights. (2) *Original DeepHit approach with J subnetworks:* We trained the DeepHit network with a separate cause-specific subnetwork per event (see Fig. 2, DeepHit^2) (3) *Single-event DNN that ignores competing events:* Similar to the first setup, we train the DeepHit network with only one subnetwork. Instead of replacing T_i by ϑ_i, we ignored the competing events and treated all individuals with an observed competing event as censored (i.e., we treated the observed time to the occurrence of the competing event as the censoring time).

Each experiment was repeated 10 times per dataset in order to reduce the effect of random sampling and random initialization on the results.

DRSA network. To assess the effectiveness of the imputation strategy on a deep neural network designed for time-to-event data analysis without competing events, we used the deep recurrent survival analysis (DRSA) architecture by Ren et al.[27]. We picked this architecture because a) it is primarily designed for a single-event discrete-time survival analysis setting and b) because DRSA differs structurally from the DeepHit architecture, therefore, allowing us to assess the effectiveness of the proposed approach with different types of deep neural networks. In contrast to DeepHit that consists of consecutive fully connected layers, DRSA consists of a layer of Long Short-Term Memory (LSTM)[33] units in addition to fully connected layers. In other words, the DRSA network consists of an initial layer that embeds the input features x_i into a set of vectors. Then through a fully connected layer, the embedded vectors are turned into a middle-representation of the input. The output of this layer is concatenated with the observed time points (t) and is fed into the recurrent layer, consisting of a series of LSTM units. In the end, a fully connected layer is used with the Sigmoid activation function to estimate the hazard rates at each time point t. For better-calibrated prediction rules and improved discriminatory power, instead of the cross-entropy loss that was used in the original DRSA network, we used the loss function derived from the negative log-likelihood of the discrete time-to-event model[16]. The loss function that was considered for the optimization consisted of two terms L_l and L_z, i.e., $\arg\min_\theta (1-\alpha)L_l(\theta) + \alpha L_z(\theta)$, where θ denotes the set of network parameters, α denotes the tuning parameter balancing the two loss terms, L_l denotes the negative log-likelihood loss and L_z denotes a part of the negative log-likelihood that was only computed for the set of uncensored instances in the training data[16].

To assess the performance of our proposed method with DRSA, we compared two different setups: (1) *New approach using DRSA with preprocessed input data:* Similar to the experiments with DeepHit, instead of the original input data, we used the modified version of the input data (with T_i replaced by ϑ_i), in which the censoring times corresponding to individuals with observed competing events were imputed using the subdistribution weights. (2) *DRSA that ignores competing events by treating them as censored:* Similar to the first setup, instead of replacing T_i by ϑ_i, we ignored the competing events by treating all individuals with an observed competing event as censored.

Again, each experiment was repeated 10 times per dataset in order to reduce the effect of random sampling and random initialization on the results.

Data description. In this subsection, we describe the datasets that were used in the experiments. To show the effectiveness of the imputation strategy, we created three sets of simulated competing risks data. Additionally, to test our method in real-world scenarios, we used two datasets from clinical and epidemiological research: The first one was collected for the CRASH-2 clinical trial[19] mentioned above; the second one was the 2013 breast cancer dataset from the Surveillance, Epidemiology, and End Results (SEER) program[34].

Simulated data. For generating simulated data, we used the discrete model by Berger et al.[35]. Their data generation approach was adopted from Fine and Gray[23] and Beyersmann et al.[36], and allowed to create datasets from a discretized subdistribution hazard model with two competing events $\epsilon_i \in \{1, 2\}$.

More specifically, Berger et al.[35] defined their discretized subdistribution hazard model based on the continuous subdistribution hazard model

$$F_1(t|\boldsymbol{x}_i) \;=\; P(T_{cont,i} \le t, \epsilon_i = 1 \,|\, \boldsymbol{x}_i) = 1 - (1 - q + q \cdot \exp(-t))^{\exp(\boldsymbol{x}_i^\top \boldsymbol{\gamma}_1)}, \tag{2}$$

where $T_{cont,i} \in \mathbb{R}^+$ denotes a continuous time variable and $\boldsymbol{\gamma}_1$ is a set of regression coefficients for individual i, with predictor variables $\boldsymbol{x}_i$. We used the parameter q to tune the probability of having the event $\epsilon_i = 1$ (defined by $P(\epsilon_i = 1|\boldsymbol{x}_i) = 1 - (1-q)^{\exp(\boldsymbol{x}_i^\top \boldsymbol{\gamma}_i)}$) and the probability of having the competing event $\epsilon_i = 2$ (defined by $P(\epsilon_i = 2|\boldsymbol{x}_i) = 1 - P(\epsilon_i = 1|\boldsymbol{x}_i) = (1-q)^{\exp(\boldsymbol{x}_i^\top \boldsymbol{\gamma}_i)}$). Further, the continuous times for the second event were drawn from an exponential model $T_{cont,i}|\epsilon_i = 2 \sim \text{Exp}(\xi_2 = \exp(\boldsymbol{x}_i^\top \boldsymbol{\gamma}_2))$, with rate ξ_2 and regression parameters $\boldsymbol{\gamma}_2$ for the predictor variables $\boldsymbol{x}_i$. To obtain grouped data, we discretized the continuous event times into $k = 20$ time-intervals using empirical quantiles. Analogous to Berger et al.[30], discrete censoring times were

Censoring rate	Type-1 rate	Type-2 rate	Training	Validation	Test
Simulated data					
47.4%	11.5%	41.1%	15,000	5000	10,000
47.6%	21.8%	30.6%	15,000	5000	10,000
48.0%	38.6%	13.4%	15,000	5000	10,000
CRASH-2 data					
16.8%	4.9%	78.3%	9729	3256	6851
16.8%	14.9%	68.3%	9729	3256	6851
SEER breast cancer data					
88.4%	6.9%	4.7%	60,898	24,361	36,539

Table 1. Characteristics of the datasets used in the experiments. The three leftmost columns represent the censoring, type-1 ($\epsilon = 1$), and type-2 ($\epsilon = 2$) rates in the training/validation/test datasets. The three rightmost columns represent the respective numbers of instances in the simulated, CRASH-2, and SEER breast cancer data. For CRASH-2, $\epsilon = 1$ indicates either death due to bleeding event (upper row) and death due to any recorded cause (lower row).

drawn from the probability distribution $P(C_i = t) = b^{(k+1-t)}/\sum_{i=1}^{k} b^i$, where the parameter $b \in \mathbb{R}^+$ affected the overall censoring rate. Furthermore, we generated four predictor variables: two of them were normally distributed, $x_1, x_2 \sim N(0, 1)$, and the other two followed a binomial distribution each, $x_3, x_4, \sim \text{Binomial}(1, 0.5)$. The regression coefficients were the same as in the work by Berger et al.[35], with $\gamma_1 = c(0.4, -0.4, 0.2, -0.2)^\intercal$ and $\gamma_2 = c(-0.4, 0.4, -0.2, 0.2)^\intercal$. We simulated datasets of size $n = 30,000$ with different type-1 event rates $q \in \{0.2, 0.4, 0.8\}$ and a *medium* censoring rate of $b = 1$. In the simulated datasets the empirical censoring rates corresponding to $b = 1$ were $\{47.4\%, 47.6\%, 48.0\%\}$, the proportion of type-1 event rates corresponding to values of q were $\{11.5\%, 21.8\%, 38.6\%\}$, and consequently type-2 event rates were $\{41.1\%, 30.6\%, 13.4\%\}$.

CRASH-2 data. The first real-world dataset used in our experiments was collected for the randomized CRASH-2 (Clinical Randomisation of an Antifibrinolyticin Significant Haemorrhage 2) trial, which was conducted in 274 hospitals in 40 countries between 2005 and 2010[19]. The data provide information on hospital death in adult trauma patients with or at risk of significant haemorrhage. Death was recorded during hospitalization of the patients for up to 28 days after randomization. Up to this date, patients had either died, been discharged alive, transferred to another hospital, or were still alive in hospital. For our analysis we used the publicly available version of the study database at https://hbiostat.org/data/. Based on Table 1 in [19], we selected eight variables for analysis: Categorical variables included the sex of the patient (male/female) and type of injury (blunt/penetrating/blunt and penetrating). Continuous and ordinal variables included total Glasgow Coma Score (range 3 to 15, median = 15), the estimated age of the patient (mean = 34.6 years, sd = 14.3 years), number of hours since injury (mean = 2.8, sd = 2.4), systolic blood pressure in mmHg (mean = 97.5, sd = 27.4), respiratory rate per minute (mean = 23.1, sd = 6.7), and heart rate per minute (mean = 104.5, sd = 21.0). After discarding patients with missing values, we analyzed this dataset in two ways: 1) We specified *death due to bleeding* as the event of interest for analysis ($\epsilon = 1$) and considered *discharge from the hospital or death due to other causes* as the competing event ($\epsilon = 2$). In this scenario, the censoring rate is 16.8%, the type-1 event rate was 4.9% and the type-2 event rate was 78.3%. 2) We specified *death from any cause* as the event for interest for analysis ($\epsilon = 1$) and considered *discharged from the hospital* as the competing event ($\epsilon = 2$). In this scenario, the censoring rate was 16.8, the type-1 event rate was 14.9% and the type-2 event rate was 68.3%. Table 1 summarizes the percentage of patients experiencing each event first. These analyses enabled us to investigate the performance of different methods for varying event rates while censoring remained the same.

SEER breast cancer data. The second real-world dataset used in our experiments was the 2013 breast cancer data from the Surveillance, Epidemiology, and End Results (SEER) program[34]. Here our focus was on female patients with breast cancer, aged 18-75 years at the time of diagnosis. We specified *patient's death due to breast cancer* as event of interest ($\epsilon = 1$) and considered *death due to other causes* as the competing event ($\epsilon = 2$). The predictor variables included TNM stage (twelve T stage and four N stage categories), tumor grade (I - IV), estrogen and progesterone receptor statuses (positive/negative), primary tumor site (nine categories), surgery of primary site (yes/no), type of radiation therapy and sequence (seven and six categories, respectively), laterality (right/left), ethnicity (white, black, American Indian/Alaska Native, Asian or Pacific Islander, unknown), Spanish origin (nine categories), and marital status at diagnosis (single, married, separated, divorced, widowed). In addition to these categorical variables, we selected the following continuous and ordinal features; patient's age at diagnosis (recorded in years, mean age = 55.6 years, standard deviation (sd) = 10.8 years), the number of positive and examined lymph nodes (0-84 and $1, 2, \ldots, 89,\ 90$, respectively), the number of primaries (1-6), and tumor size ($0, 1, \ldots, 988,\ 989$ mm). After discarding patients with missing values, $121,798$ patients remained. For this dataset the censoring rate was 88.4%, the type-1 event rate was 6.9% and the type-2 event rate was 4.7%. For a detailed explanation of the features, see the SEER text data file description at http://seer.cancer.gov.

Training setup. *Simulated data.* For our experiments we split the 30,000 instances of each set of simulated data into training ($\mathcal{D}_{train}$) , test ($\mathcal{D}_{test}$) and validation ($\mathcal{D}_{validation}$) sets randomly, making sure that the event and censoring rates were the same across the three datasets. The sizes of the train, test and validation datasets were 15,000, 10,000 and 5000 respectively. Table 1 briefly summarizes the size of the datasets used in each experiment. Since in our method the censoring times for individuals with an observed competing event are randomly imputed, we repeated the experiments 10 times and report the average performance. For each repetition, all of the individuals in training, test, and validation sets remained unchanged, except for the censoring times that were re-imputed.

CRASH-2 data. For this dataset, we used the same training setup as for the simulated data. We randomly split the 19, 836 instances into the training, test, and validation sets, using a stratified sampling approach that ensured all had approximately the same censoring and competing event rates (see Table 1). The sizes of the training, test and validation datasets were 9, 729, 6, 851 and 3, 256 respectively.

SEER data. We used the same training setup as for the other datasets. We randomly split the 121, 798 instances into the training, test, and validation sets, making sure all had 88.4%, 6.9%, and 4.7%, of censoring, event of interest and competing event rates respectively (see Table 1). The sizes of the training, test and validation datasets were 60, 898, 36, 539 and 24, 361 respectively.

Evaluation metrics. *Calibration plots based on the cumulative incidence function (CIF).* To assess the calibration of the fitted models, we performed graphical comparisons of the estimated (model-based) CIF for type-1 events and a respective nonparametric estimate obtained from the Aalen-Johansen method[37].

Specifically, for input predictor variables x_i from $\mathcal{D}_{\text{test}}$, the model-based CIF at timepoint t for the event of interest was estimated by

$$\hat{F}_1(t|x_i) = \hat{P}(T \le t, j = 1|x_i) = \sum_{s=1}^{t} \hat{P}(T = s, j = 1|x_i), \tag{3}$$

where the probability estimates $\hat{P}(\cdot)$ in (3) were taken from the output of the DeepHit network (for details, see Lee et al.[9]). Details on the Aalen-Johansen estimator, which is a covariate-free estimator of the CIF, have been given in the book by Klein et al.[37]. In our experiments, we considered a fitted DNN model to be well calibrated if the model-based and covariate-free CIF estimates agreed closely.

Concordance index (C-index[38,39]). To evaluate the discriminatory power of each method for the event of interest we used the C-index as defined by Wolbers et al.[40]. For a pair of independent individuals i and j in the $\mathscr{D}_{\text{test}}$, this measure compares the ranking of a *risk marker* $M(t, x_i)$ at timepoint t with the ranking of the survival times of the event of interest. More specifically, summarizing all competing events by $\epsilon = 2$, the C-index is defined by

$$C_1(t) := P\big(M(t, x_i) > M(t, x_j) \,|\, \epsilon_i = 1 \text{ and } T_i \le t \text{ and } (T_i < T_j \text{ or } \epsilon_j = 2)\big). \tag{4}$$

In our experiments we defined $M(t, x)$ by the cumulative incidence function (Equation 3). Ideally, the C-index takes value 1 if the rankings of the risk marker and the type-1 survival times are in perfect disagreement (i.e., larger marker values are associated with smaller survival times). For our experiments, we used the inverse-probability-weighted estimator by Wolbers et al.[40] (Equation 4) that is implemented in the R package **pec**.

Results

The calibration plots for the various model fits are presented in Fig. 3. It is seen that despite the smaller learning capacity of the imputation-based DeepHit[1] approach, this network resulted in similarly well-calibrated models as the DeepHit[2] with two sub-networks. Note that in all cases, using the sub-distribution weights for imputing the censoring times led to a better calibration compared to the single-event DeepHit architecture that treated individuals with an observed competing event as censored (thus ignoring the competing events).

Generally, the calibration of the overall average CIF estimate improved with our method when the rate type-1 events became larger. This is seen from the last row of Fig. 3. For the same censoring rates and predictor variables (for CRASH-2), DeepHit[2] resulted in an underestimation of the CIF when the rate of type-1 events was high. This is also evident in the results from our experiments on simulated data. On the other hand, our proposed method showed an overall less sensitivity to the type-1 event rate. This effect could possibly be due overfitting issues, as adding an additional sub-network for each competing event to the architecture increases the learning capacity of the network without providing enough data to train each pathway.

The calibration plots for training with DRSA are presented in Fig. 4. It is seen that despite the single-event structure of the DRSA, this network resulted in a well-calibrated model when the type-1 event rate was small. In all cases, using the sub-distribution weights for imputing the censoring times led to a better calibration compared to the experiments that treated individuals with an observed competing event as censored (thus ignoring the competing events). For the same censoring rates and predictive variables, DRSA resulted in an underestimation of the CIF when the rate of type-1 events was high. On the other hand, again our proposed method showed an overall less sensitivity to the type-1 event rate compared to when the competing event was ignored.

Analogous to the results from the calibration plots, the C-indices obtained from our imputation-based method showed a discriminatory power that was similar to the respective performance of the other methods (see Table 2). In a number of settings, the discriminatory power even improved when using our method. For instance, in the experiments with the simulated data, the estimated mean C-index was highest for the DeepHit[1] method with imputed censoring times. For CRASH-2 with a type-1 event rate of 4.9% the observed difference (0.01%)

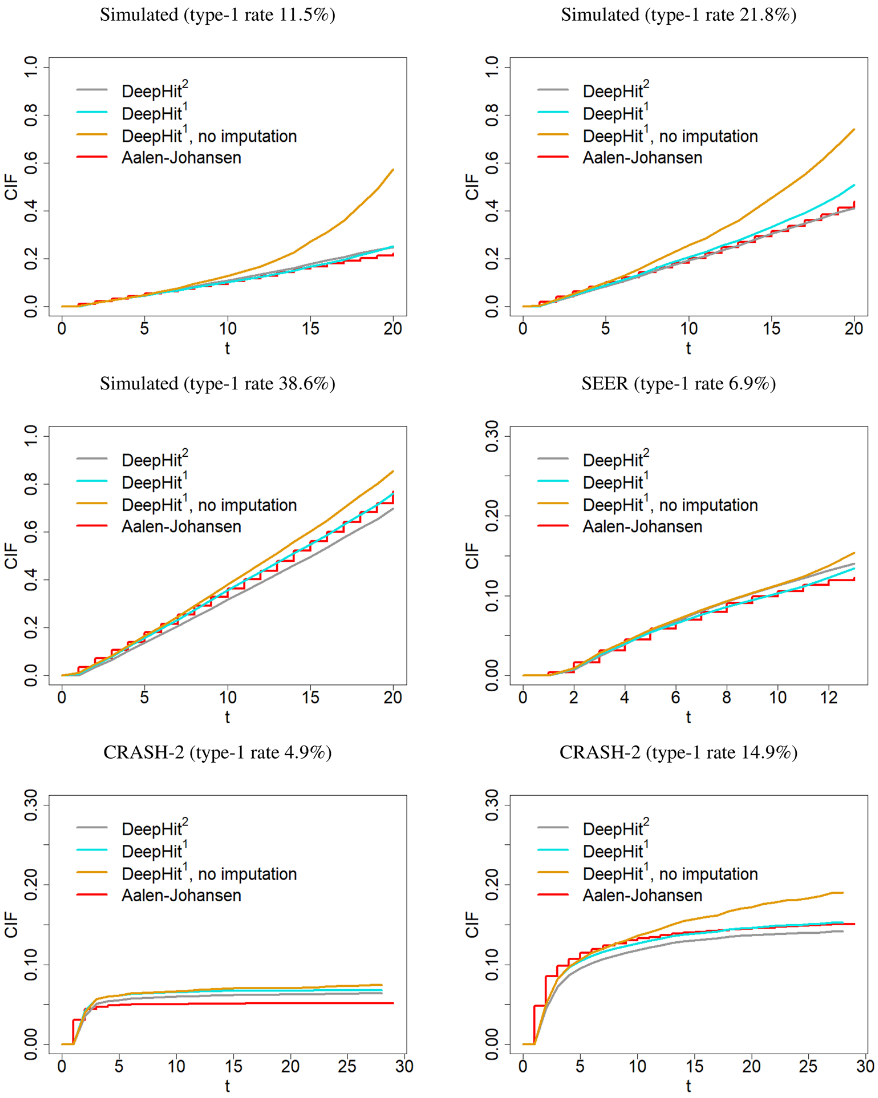

Figure 3. Calibration plots obtained from the test data in Table 1, using the DeepHit architecture. Each plot presents the averaged type-1 cumulative incidence functions as obtained from (i) training the DeepHit^1 with the preprocessed data (cyan), (ii) training DeepHit^1 treating individuals with observed competing events as censored (orange), and (iii) training DeepHit^2 for both the event of interest and the competing event (gray). Red curves refer to the nonparametric Aalen-Johansen reference curves.

between imputation-based DeepHit^1 and DeepHit^2 was small. For the type-1 event rate of 14.9% our proposed method performed slightly better. For the SEER breast cancer data, however, DeepHit^1 without imputation had the best average performance with regard to the *C*-index. This could be due to the fact that the rate of observed competing events was low to the degree that treating the respective event times as censoring times might not have substantially affected the censoring survival function.

Analogous to the experiments with DeepHit, for DRSA, the *C*-indices obtained from our imputation-based method showed an improved discriminatory power compared to the scenario when competing event time was used as censoring (see Table 3). It can be observed that the gap between the performance of our imputation method and ignoring the competing events became smaller with the decrease of type-2 event rate. The reason could be that by the decrease of the observed competing events rate, treating the respective event times as censoring times might not have substantially affected the censoring survival function. Overall, compared to DRSA, DeepHit showed better discriminatory power on the simulated data. Note, however, that systematic performance comparisons of different deep survival architectures are beyond the scope of this work.

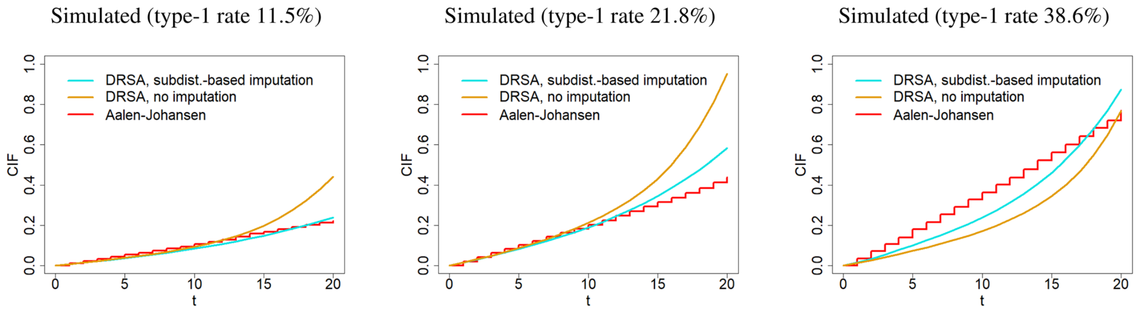

Figure 4. Calibration plots obtained from the simulated test data in Table 1 using the DRSA architecture. Each plot presents the averaged type-1 cumulative incidence functions as obtained from (i) training the DRSA network with the preprocessed training data (cyan) and (ii) training DRSA treating individuals with observed competing events as censored (orange). Red curves refer to the nonparametric Aalen-Johansen reference curves.

Data	Type-1-rate	Type-2-rate	DeepHit[1]	DeepHit[1], no imp.	DeepHit[2]
CRASH-2	4.9%	78.3%	78.17 ± 1.04	76.80 ± 4.96	**78.18** ± 0.94
CRASH-2	14.9%	68.3%	**80.14** ± 1.77	79.88 ± 2.01	80.05 ± 4.23
SEER	6.9%	4.7%	81.75 ± 3.46	**81.80** ± 3.49	81.73 ± 3.34
Simulated	11.5%	41.1%	**64.13** ± 0.75	62.58 ± 2.17	63.71 ± 0.96
Simulated	21.8%	30.6%	**65.90** ± 0.69	64.59 ± 2.25	65.20 ± 3.26
Simulated	38.6%	13.4%	**66.05** ± 0.47	64.97 ± 2.51	64.39 ± 6.26

Table 2. Mean estimated *C*-indices (averaged over time) with estimated standard deviations, as obtained from training the DeepHit architecture on the simulated, CRASH-2, and SEER breast cancer data. DeepHit[1] = DeepHit architecture with one sub-network trained with the preprocessed input data; DeepHit[2] = DeepHit architecture with two subnetworks; DeepHit[1], no imp. = DeepHit architecture with one sub-network trained on the original input data (treating individuals with observed competing events as censored individuals). Best-performing methods are marked bold. Note that the *C*-indices must be compared within each row, as the datasets used for training were different in terms of size, censoring, and event rates across the rows. For CRASH-2, in the upper and the lower rows $\epsilon = 1$ indicates death due to bleeding and death due to any recorded cause, respectively. The numbers in this table are obtained from the test datasets.

Data	Type-1-rate	Type-2-rate	DRSA, subdist.-based imp.	DRSA, no imp.
Simulated	11.5%	41.1%	**58.04** ± 0.88	55.62 ± 0.86
Simulated	21.8%	30.6%	**60.10** ± 0.95	57.60 ± 0.93
Simulated	38.6%	13.4%	**64.29** ± 0.93	63.41 ± 1.00

Table 3. Mean estimated *C*-indices (averaged over time) with estimated standard deviations, as obtained from training the DRSA architecture on the simulated data. The first column on the right-hand side contains results from DRSA architecture trained with the preprocessed input data; The second column shows the results from the DRSA architecture, trained on the original input data (treating individuals with observed competing events as censored individuals). Best-performing methods are marked bold. Note that the *C*-indices must be compared within each row, as the datasets used for training are different in terms of censoring and event rates across the rows. The numbers in this table are obtained from the test datasets.

In terms of execution time, we observed that the average time needed for training the deep networks reduced by 21% for the simulated data, 10% for the SEER, and 37% for the CRASH-2 dataset using our method. This time reduction is possibly due to the reduced number of parameters involved in the training of DeepHit[1] compared to DeepHit[2] (see Table 4). Consequently, in applications with more than one competing event, where three or more subnetworks are added to the architecture, the decrease in computation time when using our algorithm is expected to be even greater. The average number of iterations, however, was on the same order of magnitude for both DeepHit[1] and DeepHit[2]. For all datasets on average DeepHit[1] took 15, 022 iterations and DeepHit[2] 15, 277. Note that the stopping criterion for all of the networks was the performance on the validation data.

	Simulated	SEER	CRASH-2
	Time | #itr	Time | #itr	Time | #itr
DeepHit^1	184.78 | 10, 666	827.97 | 22, 600	116.47 | 11, 800
DeepHit^2	235.32 | 9133	918.39 | 22, 300	185.30 | 14, 400

Table 4. Average time (in seconds) and number of iterations needed for training DeepHit^1 and DeepHit^2 per dataset. Performance on validation data was used as the stopping criterion.

Discussion

Even though deep neural networks are increasingly used for survival analysis, it is still relatively complicated to adapt the available methodology to situations with competing events. This is in contrast to the classical statistical literature, in which a wide variety of methods are available[20–23,41], and in which it is widely agreed that competing-risks analyses are often necessary to avoid biased estimation results and/or predictions[36]. Although several adaptations to DNN architectures have been proposed recently[9,11,24], these adaptions rely on a huge number of parameters, making network training and regularization a challenging task. In this work, we showed that an imputation strategy based on subdistribution weights could convert the competing risks survival data into a dataset that is specifically tailored to analyzing the event of interest only. This conversion enables the use of any of the much simpler deep survival network architectures that are designed to handle a single event of interest in the presence of right censoring. Our experiments on simulated and real-world datasets illustrated that this preprocessing step not only simplifies the training in terms of number of parameters and running time but also preservers the accuracy in terms of discriminatory power and calibration. The method could be further stabilized by implementing a multiple imputation approach (analogous to the continuous-time method by Ruan and Gray[32]); however, such an approach would dramatically increase the run time and would be infeasible in the context of training DNN architectures. Further, in our experiments we observed that multiple imputations did not have a major effect on predictive performance in our datasets containing several thousands of instances with event rates larger than $\sim$ 5%. Our codes for simulated data generation, censoring time imputation, and the experiments are available at https://github.com/shekoufeh/Deep-Survival-Analysis-With-Competing-Events.

Received: 16 September 2021; Accepted: 24 February 2022
Published online: 09 March 2022

References

1. Ren, S., He, K., Girshick, R. & Sun, J. Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp. 91–99 (2015).
2. Karpathy, A. & Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3128–3137 (2015).
3. Affonso, C. *et al.* Deep learning for biological image classification. *Exp. Syst. Appl.* **85**, 114–122 (2017).
4. Abdel-Zaher, A. M. & Eldeib, A. M. Breast cancer classification using deep belief networks. *Exp. Syst. Appl.* **46**, 139–144 (2016).
5. Graves, A., Mohamed, A. & Hinton, G. Speech recognition with deep recurrent neural networks. In: Proceedings of the 2013 IEEE international conference on acoustics, speech and signal processing, pp. 6645–6649 (IEEE, New York, 2013).
6. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, Cambridge, MA, 2016).
7. Sonabend, R., Király, F.J., Bender, A., Bischl, B. & Lang, M. mlr3proba: an r package for machine learning in survival analysis. arXiv preprint arXiv:2008.08080 (2020).
8. Giunchiglia, E., Nemchenko, A. & van der Schaar, M. RNN-SURV: a deep recurrent model for survival analysis. In: Proceedings of the 27th International Conference on Artificial Neural Networks, pp. 23–32 (Springer, Cham, 2018).
9. Lee, C., Zame, W. R., Yoon, J. & van der Schaar, M. DeepHit: A deep learning approach to survival analysis with competing risks. In: Proceedings of the thirty-second AAAI conference on artificial intelligence, pp. 2314–2321 (AAAI Press, Palo Alto, 2018).
10. Nezhad, M. Z., Sadati, N., Yang, K. & Zhu, D. A deep active survival analysis approach for precision treatment recommendations: application of prostate cancer. *Exp. Syst. Appl.* **115**, 16–26 (2019).
11. Gupta, G., Sunder, V., Prasad, R. & Shroff, G. Cresa: a deep learning approach to competing risks, recurrent event survival analysis. In: Pacific-Asia conference on knowledge discovery and data mining, pp. 108–122 (Springer, 2019).
12. Kvamme, H. & Borgan, Ø. Continuous and discrete-time survival prediction with neural networks. arXiv preprint arXiv:1910.06724 (2019).
13. Zhu, X., Yao, J., Zhu, F. & Huang, J. Wsisa: making survival prediction from whole slide histopathological images. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7234–7242 (2017).
14. Zhu, X., Yao, J. & Huang, J. Deep convolutional neural network for survival analysis with pathological images. In: Proceedings of the 2016 IEEE international conference on bioinformatics and biomedicine (BIBM), pp. 544–547 (IEEE, New York, 2016).
15. Ren, K., et al. Deep recurrent survival analysis. In: Proceedings of the thirty-third AAAI conference on artificial intelligence, pp. 4798–4805 (AAAI Press, Palo Alto, 2019).
16. Gorgi Zadeh, S. & Schmid, M. Bias in cross-entropy-based training of deep survival networks. IEEE Trans. Pattern Anal. Mach. Intell. (2020).
17. Faraggi, D. & Simon, R. A neural network model for survival data. *Stat. Med.* **14**, 73–82 (1995).
18. Lee, M.-L.T. & Whitmore, G.A. Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. Stat. Sci. 501–513 (2006).
19. CRASH-2 Trial Collaborators. Effects of tranexamic acid on death, vascular occlusive events, and blood transfusion in trauma patients with significant haemorrhage (CRASH-2): a randomised, placebo-controlled trial. Lancet **376**, 23–32 (2010).
20. Lau, B., Cole, S. R. & Gange, S. J. Competing risk regression models for epidemiologic data. *Am. J. Epidemiol.* **170**, 244–256 (2009).
21. Austin, P. C., Lee, D. S. & Fine, J. P. Introduction to the analysis of survival data in the presence of competing risks. *Circulation* **133**, 601–609 (2016).
22. Prentice, R.L., et al. The analysis of failure times in the presence of competing risks. Biometrics 541–554 (1978).

23. Fine, J. P. & Gray, R. J. A proportional hazards model for the subdistribution of a competing risk. *J. Am. Stat. Assoc.* **94**, 496–509 (1999).
24. Nagpal, C., Li, X. R. & Dubrawski, A. Deep survival machines: fully parametric survival regression and representation learning for censored data with competing risks. IEEE J. Biomed. Health Inform. (2021).
25. Katzman, J. L. *et al.* Deep survival: a deep cox proportional hazards network. *Statistics* **1050**, 1–10 (2016).
26. Yousefi, S. *et al.* Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Sci. Rep.* **7**, 1–11 (2017).
27. Ren, K. *et al.* Deep recurrent survival analysis. *Proc. AAAI Conf. Artif. Intell.* **33**, 4798–4805 (2019).
28. Kleinbaum, D. G. & Klein, M. *Survival analysis* (Springer, New York, 2010).
29. Schmid, M. & Berger, M. Competing risks analysis for discrete time-to-event data. Wiley interdisciplinary reviews: computational statistics e1529 (2020).
30. Berger, M., Schmid, M., Welchowski, T., Schmitz-Valckenberg, S. & Beyersmann, J. Subdistribution hazard models for competing risks in discrete time. *Biostatistics* **21**, 449–466 (2020).
31. Thomas Welchowski and Matthias Schmid. R: Discrete Time Survival Analysis (2019).
32. Ruan, P. K. & Gray, R. J. Analyses of cumulative incidence functions via non-parametric multiple imputation. *Stat. Med.* **27**, 5709–5724 (2008).
33. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
34. National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch. The surveillance, epidemiology and end results (SEER) research data. Cases diagnosed in 1973–2010, follow up cutoff Dec 2010, released on April 2013, based on the November 2012 submission https://seer.cancer.gov/ (2013).
35. Berger, M. & Schmid, M. Semiparametric regression for discrete time-to-event data. *Stat. Model.* **18**, 1–24 (2018).
36. Beyersmann, J., Allignol, A. & Schumacher, M. *Competing risks and multistate models with R* (Springer, New York, 2011).
37. Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G. & Scheike, T. H. *Handbook of survival analysis* (CRC Press, Boca Raton, 2016).
38. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *J. Am. Med. Assoc.* **247**, 2543–2546 (1982).
39. Harrell, F. E., Lee, K. L., Califf, R. M., Pryor, D. B. & Rosati, R. A. Regression modeling strategies for improved prognostic prediction. *Stat. Med.* **3**, 143–152 (1984).
40. Wolbers, M., Blanche, P., Koller, M. T., Witteman, J. C. & Gerds, T. A. Concordance for prognostic models with competing risks. *Biostatistics* **15**, 526–539 (2014).
41. Wolbers, M., Koller, M.T., Witteman, J.C. & Steyerberg, E.W. Prognostic models with competing risks: methods and application to coronary risk prediction. Epidemiology 555–561 (2009).

Acknowledgements

The authors thank Dr. Lothar Haberle (Department of Gynecology, Obstetrics and Mammology, University Hospital Erlangen, Germany) for valuable comments on the analysis of the SEER data. Support from the German Research Foundation (Grant SCHM 2966/2-1), is gratefully acknowledged.

Author contributions

M.S. and S.G. conceived the methodology with inputs from C.B.. S.G. conceived and conducted the experiments, and prepared Figs. 2, 3, and 4. S.G. and M.S. wrote the manuscript, with contributions from C.B. C.B. performed the preprocessing of datasets and prepared Fig. 1. All authors analyzed the results and reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.G.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

3.4 Publication D: Random Survival Forests with Competing Events: A Subdistribution-Based Imputation Approach

Behning C, Bigerl A, Wright MN, Sekula P, Berger M, Schmid M. Random Survival Forests With Competing Events: A Subdistribution-Based Imputation Approach. In: Biometrical Journal 2024; 66 (6): e202400014

`https://doi.org/10.1002/bimj.202400014`

Implementations and evaluation code are available on GitHub:

`https://github.com/cbehning/ranger/tree/competing_risks_subdist`

and

`https://github.com/cbehning/rsf_competing_events`

Biometrical Journal

WILEY-VCH

Biometrical Journal

RESEARCH ARTICLE OPEN ACCESS

Random Survival Forests With Competing Events: A Subdistribution-Based Imputation Approach

Charlotte Behning[1] | Alexander Bigerl[2] | Marvin N. Wright[3,4,5] | Peggy Sekula[6] | Moritz Berger[1] | Matthias Schmid[1]

[1]Institute of Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Bonn, Germany | [2]DICE Group, Department of Computer Science, Paderborn University, Paderborn, Germany | [3]Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany | [4]Faculty of Mathematics and Computer Science, University of Bremen, Bremen, Germany | [5]Section of Biostatistics, Department of Public Health, University of Copenhagen, Copenhagen, Denmark | [6]Institute of Genetic Epidemiology, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

Correspondence: Charlotte Behning (behning@imbie.uni-bonn.de)

Received: 15 January 2024 | **Revised:** 3 July 2024 | **Accepted:** 10 July 2024

Funding: M.N.W. was supported by the German Research Foundation (DFG) - Emmy Noether Grant 437611051. The German Chronic Kidney Disease study was funded by grants from the German Federal Ministry of Education and Research (BMBF, grant number 01ER0804), the KfH Foundation for Preventive Medicine, and corporate sponsors (http://www.gckd.org).

Keywords: competing events | discrete time-to-event data | imputation | random survival forest | subdistribution hazard

ABSTRACT

Random survival forests (RSF) can be applied to many time-to-event research questions and are particularly useful in situations where the relationship between the independent variables and the event of interest is rather complex. However, in many clinical settings, the occurrence of the event of interest is affected by competing events, which means that a patient can experience an outcome other than the event of interest. Neglecting the competing event (i.e., regarding competing events as censoring) will typically result in biased estimates of the cumulative incidence function (CIF). A popular approach for competing events is Fine and Gray's subdistribution hazard model, which directly estimates the CIF by fitting a single-event model defined on a subdistribution timescale. Here, we integrate concepts from the subdistribution hazard modeling approach into the RSF. We develop several imputation strategies that use weights as in a discrete-time subdistribution hazard model to impute censoring times in cases where a competing event is observed. Our simulations show that the CIF is well estimated if the imputation already takes place outside the forest on the overall dataset. Especially in settings with a low rate of the event of interest or a high censoring rate, competing events must not be neglected, that is, treated as censoring. When applied to a real-world epidemiological dataset on chronic kidney disease, the imputation approach resulted in highly plausible predictor–response relationships and CIF estimates of renal events.

1 | Introduction

Survival analysis aims to model the time until the occurrence of a specific event (e.g., progression or death due to a certain disease) in dependence on a set of covariates. In clinical contexts, time-to-event data are often collected in observational studies that are prone to right censoring. Right censoring happens, for example, when patients drop out of a study or do not experience their event before the end of the observation period. In addition to a single *event of interest*, other event types are often recorded in observational studies and present in survival datasets. Often, the occurrence of these *competing events* cannot be assumed to be independent of the occurrence of the event of interest, especially if shared underlying (disease) mechanisms or shared risk factors are present.

An example would be examining kidney failure (KF) as the event of interest in patients with chronic kidney disease (CKD), while death by other causes than KF is a competing event (Hsu et al. 2017). In the German Chronic Kidney Disease (GCKD) study (Titze et al. 2015), for instance, 5217 participants with CKD are followed up annually, so data can be evaluated at the discrete time points corresponding to 1-year time intervals. One of the aims of the study is to better understand the factors underlying the progression of the disease. Potential risk factors that were collected in the study at baseline included, for example, leading kidney disease, as well as kidney function measures, such as serum creatinine, estimated glomerular filtration rate (eGFR), and U-albumin/creatinine ratio (UACR). Since CKD is a risk factor for heart failure (HF) and CKD and HF share common risk factors (Beck et al. 2015), death (e.g., from cardiovascular causes) should be considered as a competing event.

A popular approach to analyzing survival data (i.e., time to first event) in the presence of competing events is the subdistribution hazard model by Fine and Gray (1999), which extends the classical Cox proportional hazard model (Cox 1972). The Fine and Gray model introduces a subdistribution hazard function, which is a modification of the hazard function in traditional survival analysis. This function quantifies the instantaneous rate of the event of interest occurring, given that the subject has not yet experienced the event of interest until that time (assuming that the event of interest will never occur first once a competing event has already occurred (cf. Fine and Gray 1999).

As with other classical regression approaches, the subdistribution hazard model is not designed for high-dimensional data settings or complex covariate–risk relationships. In such scenarios, machine learning models such as deep survival neural networks (e.g., Giunchiglia, Nemchenko, and van der Schaar 2018; Gupta et al. 2019; Lee et al. 2018) and random survival forests (RSF) can be applied (Ishwaran et al. 2008; Schmid, Wright, and Ziegler 2016; Wright, Dankowski, and Ziegler 2017). While neural networks can be most beneficial for unstructured data, such as text and images, random forests might be advantageous for structured data exploration and for identifying important clinical covariates (Archer and Kimes 2008). Also, random forests are easy to train and require less resource-consuming hyperparameter tuning.

While numerous methods for competing events exist in classical regression, only some implementations of machine learning models for survival analysis consider competing events. In existing approaches, competing events in random forests are addressed by, for example, adapting the split rules (Ishwaran et al. 2014; Therrien and Cao 2022) or by using pseudo-value regression approaches (Mogensen and Gerds 2013). The latter method transforms the categorical event status into a continuous pseudo-value. Consequently, a random forest with regression trees is fitted instead of survival trees.

In this paper, we take a different approach for modeling competing risk data with RSF: Rather than introducing new split rules or new architectures for competing events, we transform the competing event problem into a single-event problem. This is achieved by manipulating the (input) dataset via an appropriately defined imputation scheme. More specifically, we consider three types of imputation approaches: In the first approach, the dataset is only preprocessed once before training the RSF. In the other two approaches, the dataset is adjusted directly at the tree instance of the forest: at the root node of the trees or at every node of the trees. As a consequence, well-established split rules and variable importance measures of single-event RSF can be applied. Also, the cumulative incidence function (CIF) for the event of interest can be directly calculated from the output of the single-event RSF.

The idea of using imputed censoring times instead of the observed competing event time has been applied successfully already for classical statistical modeling and neural networks: Ruan and Gray (2008) presented an imputation approach for continuous-time and semiparametric models based on Kaplan–Meier estimates. Gorgi Zadeh, Behning, and Schmid (2022) took a similar approach and proposed a method to train single-event deep neural survival networks on competing-event data, in which the unobserved censoring times of subjects with a competing event were imputed using subdistribution weights.

In this article, we describe the proposed methods and use a simulation study to evaluate their applicability and performance metrics in different situations. Finally, we report on a first application of the methods to real data obtained in the GCKD study.

2 | Methods

2.1 | Discrete Survival Analysis for Competing Risks

The aim of our proposed method is to estimate the CIF for an event of interest given a set of covariates. In a typical setting with right-censored data, we assume to follow-up the subjects $i = 1, \dots, n$ with baseline covariates $X_i = (x_{i1}, \dots, x_{ip})^T$. Either an event time T_i or a censoring time C_i is observed, with the status indicator $\Delta_i = I(T_i \leq C_i)$ and the type of event denoted by e_i. For each subject, either the event of interest ($e_i = 1$, $\Delta_i = 1$), a competing event ($e_i \neq 1$, $\Delta_i = 1$) or a censoring event ($\Delta_i = 0$) is observed. Just as in Fine and Gray's modeling approach, all competing events $e_i > 1$ are combined into one single competing event, denoted $e_i = 2$. We assume that the event time T_i and the censoring time C_i are independent random variables (random censoring). In a naive approach, where competing events are ignored and treated as censored, the random censoring assumption may be violated. We further assume that the censoring mechanism is noninformative, meaning that the distributions of T_i and C_i do not share any common parameters. In our approach, time is modeled on a discrete scale (possibly after grouping the continuous times into intervals), that is, $T_i \in \{1, 2, \dots, k\}$, where k denotes the maximum observable time (interval). This is motivated by the observation that most versions of RSF implicitly treat time as an ordinal variable (Ishwaran et al. 2008), and that many other available implementations of machine learning methods also use discrete-time data structures (e.g., Ren et al. 2019).

In this article, we focus on modeling the occurrence of the event of interest ($e_i = 1$). The CIF for the event of interest is defined as $F_1(t|X_i) = P(T_i \leq t, e_i = 1|X_i)$, so the probability of experiencing the event of interest at time t or prior with a given set of covariates X_i.

2.2 | Random Survival Forest for Single Events

The central architecture of the RSF is similar to the standard random forest approach (Breiman 2001; Ishwaran et al. 2008). In the first step, a number of (bootstrap) samples are generated. Next, a survival tree (Hothorn et al. 2004) is grown on each bootstrap sample. At each tree, *mtry* covariates are considered for splitting into child nodes, and the best split is selected. Many split rules have been proposed, including splitting based on the maximum log-rank statistic, C-index, Hellinger distance, and many more (Schmid et al. 2020). In this paper, we use the log-rank statistic, which can deal with both continuous and discrete-time data. The tree is grown until it reaches a termination constraint, for example, tree depth, minimum number of observations, or if no increase with respect to splitting criteria is possible. A cumulative hazard function (CHF; $H_1(t|X_i)$) is calculated at the terminal nodes of each tree. Averaging across all trees leads to the ensemble CHF. In settings without competing events, the CIF can be obtained from the CHF by $F_1(t|X_i) = 1 - \exp(-H_1(t|X_i))$.

2.3 | Imputation Using Subdistribution Weights

To enable the algorithm to use split rules designed for single-event scenarios, we propose first to impute censoring times in case competing events were observed. For this, we estimate the subdistribution weights based on the censoring mechanism in the dataset. The subdistribution weights for subjects who experience a competing event are defined as in Berger et al. (2020):

$$w_{it} := \frac{\hat{G}(t-1)}{\hat{G}(\tilde{T}_i - 1)}, \quad \tilde{T}_i < t \le k-1, \quad \tilde{T}_i = \min(T_i, C_i),$$

for all time points t after the observed competing event time. Here, $\hat{G}(t)$ is an estimate of the censoring survival function $G(t) = P(C_i > t)$. Based on the subdistribution weights, we sample a censoring time with probability $P(\hat{C}_i = t) = \Delta w_{it} = w_{it-1} - w_{it}$. Thus, the imputation changes the data as follows: For subjects experiencing a competing event, the competing event time $T_{e_i=2}$ is replaced by the estimated censoring time $\hat{C}_i$. The observed times $T_{e_i=1}$ or C_i remain unchanged for subjects with an event of interest or a censoring recorded. The imputed data are then used as input data for a single-event RSF, and estimates of the CIF are obtained as described in the previous subsection.

The RSF architecture allows the introduction of the described imputation at several stages of the fitting procedure. We propose the following three options:

1. Single imputation of the entire (training) dataset, performed outside the RSF architecture.
2. Imputation in the root node of each tree in the dataset. With this approach, the weights are calculated on the subset of data in the respective tree only.
3. Imputation in each node of each tree. Here, the weights are calculated only on the samples present in the respective node.

To gain an understanding of the distribution of the true C_i compared to the imputed $\hat{C}_i$, or the resulting $\hat{G}(t)$, for the three imputation approaches, please see the Illustration subsection below.

2.3.1 | Implementation

We incorporated the described imputation approaches in the C++ implementation of the R package **ranger** (Wright and Ziegler 2017). The implementation involved adding a function to the survival trees that calculates a life table estimate of the censoring survival function $\hat{G}(t)$ analogous to the function `estSurvCens` of the R package **discSurv** (Welchowski et al. 2022). The C++ command line interface has been used for benchmarks described below. The source code can be found here https://github.com/cbehning/ranger.

2.4 | Simulation Setup

We conducted a simulation study to investigate whether subdistribution-based imputation in the case of competing events can improve the estimation of the CIF in RSF compared to ignoring the competing events.

2.4.1 | Data-Generating Mechanisms

In each simulation run, we created a set of subjects $i = 1, \dots, n$, with $n = 1000$. For each subject, we first generated a vector of 50 normally distributed covariates $X_1, \dots, X_{50} \sim \mathcal{N}(0,1)$. Next, three time variables were created: a time $T_{e_i=1}$ for the event of interest, a time $T_{e_i=2}$ for the competing event, and a censoring time C_i. Afterwards, we sampled from a binary distribution with parameter $q \in (0,1)$ whether the event of interest ($e_i = 1$) or the competing event ($e_i = 2$) was observed (see below). Next, the status indicator Δ_i was generated as follows: the subject was censored if the censoring time was before the event time ($\Delta_i = 0$). If the censoring time for this subject was after the event time, the subject remained uncensored ($\Delta_i = 1$).

The experimental design used by Beyersmann, Allignol, and Schumacher (2011) and Berger et al. (2020) was adapted to create the event times and the censoring times. They simulated the event times $T_{e_i=1}$ based on a time-continuous subdistribution hazard model defined by

$$\begin{aligned} F_1(t|X_i) &= P(T_{cont,i} \le t, e_i = 1 \,|\, X_i) \\ &= 1 - (1 - q + q \cdot \exp(-t))^{\exp(\eta_1(X_i))}, \end{aligned}$$

where $T_{cont,i}$ was a true underlying continuous time variable for the event of interest and $\eta_1(X_i)$ was a linear predictor associated with the subdistribution time, which is described in more detail below. The parameter q was associated with the rate of the event of interest by $P(e_i = 1|X_i) = 1 - (1-q)^{\exp(\eta_1(X_i))}$. The continuous event times for competing events were drawn from an exponential distribution with

$$T_{cont,i} | e_i = 2 \sim \text{Exp}(\lambda = \exp(\eta_2(X_i))),$$

where $\eta_2(X_i)$ is a linear predictor associated with the competing event time.

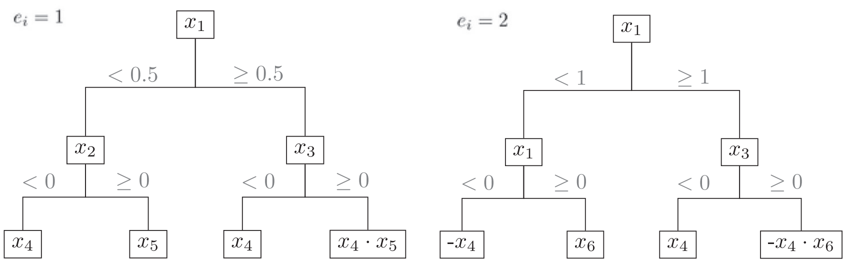

FIGURE 1 | Specification of the covariate–risk relationships in the simulation Setup 1 for the event of interest (left) and the competing event (right).

To discretize the continuous event times, we categorized the event times into $k = 20$ intervals with interval borders obtained by empirical quantiles of width 5%. The empirical quantiles were pre-estimated once per parameter q from an independent sample with 1,000,000 observations.

The discrete censoring times were generated from

$$P(C_i = t) = b^{(k+1-t)} \bigg/ \sum_{j=1}^{k} b^j,$$

where the parameter b was associated with the overall censoring rates. As in Berger et al. (2020), the parameter q was set to $q \in \{0.2, 0.4, 0.8\}$ and the parameter $b \in \{0.85, 1, 1.25\}$, corresponding to low, medium, and high censoring rates of $\{24\%, 47\%, 76\%\}$ (see Figures S1 and S2).

The following two covariate–risk relationships were investigated in this simulation study.

2.4.2 | Setup 1: Tree-Like Covariate–Risk Relationship

To mimic a rather complex relationship between covariates and event times, we modified the linear predictor functions used in Berger et al. (2020) to have a tree-like structure as depicted in Figure 1. The covariates X_1, X_2, X_3, X_4, X_5 are associated with the event of interest, and the covariates X_1, X_3, X_4, X_6 are associated with the competing event. The tree-like predictor function for the event of interest can be written as follows:

$$\begin{aligned}\eta_1(X_i) = {} & I(X_{i1} < 0.5) \cdot (I(X_{i2} < 0) \cdot X_{i4} + I(X_{i2} \geq 0) \cdot X_{i5}) \\ & + I(X_{i1} \geq 0.5) \cdot (I(X_{i3} < 0) \cdot X_{i4} + I(X_{i3} \geq 0) \cdot X_{i4} \cdot X_{i5}),\end{aligned}$$

where $I(\cdot)$ is the indicator function. The predictor for the competing event is given by

$$\begin{aligned}\eta_2(X_i) = {} & I(X_{i1} < 1) \cdot (I(X_{i1} < 0) \cdot (-X_{i4}) + I(X_{i1} \geq 0) \cdot X_{i5}) \\ & + I(X_{i1} \geq 1) \cdot (I(X_{i3} < 0) \cdot X_{i4} + I(X_{i3} \geq 0) \cdot X_{i4} \cdot X_{i6}).\end{aligned}$$

2.4.3 | Setup 2: Interactions

In a second simulation setting, multiple interaction terms are included in the data-generating model. Here, the predictors for the event of interest e_1 and the competing event e_2 are specified as follows:

$$\begin{aligned}\eta_1(X_i) = {} & 2 \cdot (X_{i1} \cdot X_{i2} \cdot X_{i3} + X_{i1} \cdot X_{i4} \cdot X_{i5} + X_{i1} \cdot X_{i3} \cdot X_{i5} \\ & + X_{i1} \cdot X_{i3} \cdot X_{i4} + X_{i2} \cdot X_{i3} \cdot X_{i4}),\end{aligned}$$

$$\begin{aligned}\eta_2(X_i) = {} & 2 \cdot (X_{i1} \cdot X_{i3} + X_{i4} \cdot X_{i6} \cdot X_{i7} + X_{i1} \cdot X_{i4} \cdot X_{i6} \\ & + X_{i1} \cdot X_{i3} \cdot X_{i7} + X_{i1} \cdot X_{i3} \cdot X_{i4}).\end{aligned}$$

In this setup, the covariates X_1, X_3, and X_4 are associated with both events, while X_2 and X_5 are only associated with the event of interest and X_6 and X_7 are only associated with the competing event. Only the interaction term $X_1 \cdot X_3 \cdot X_4$ is shared between both linear predictors. Thus, the dependency structure is similar to Setup 1, but here X_7 is added.

As illustrated in Table 1, the simulated datasets included event times for the event of interest as well as the competing event and censoring times. The competing event times need to be replaced by the (true or estimated) censoring times to make the simulated competing event datasets usable in the single-event RSF. After replacement, the status for the subjects with competing event was set to "censored" ($\Delta_i = 0$). Table 2 illustrates the different imputation strategies for obtaining a reference dataset (A), a dataset preprocessed outside the RSF (B, C), and a dataset processed within the RSF (D). More specifically, the following imputation methods to estimate the CIF were compared:

1. *Reference:* If a subject i experiences the competing event, this is replaced with the true (simulated) censoring time C_i in the dataset. With these input data, the RSF for single events will model the true censoring rate and serves as a reference (see Table 2A).
2. *Naive approach:* Ignoring the competing event and treating the competing event time $T_{e_i=2}$ as if a censoring happened (see Table 2B).
3. *Impute once (imputeOnce):* Single imputed dataset before fitting the standard single-event RSF implementation (see Table 2C).
4. *Impute in root (imputeRoot):* RSF implementation with imputation in each root node, thus imputing once in each tree on all subjects available at the tree's root node (see Table 2D).
5. *Impute in each node (imputeNode):* RSF implementation with imputation in every node, thus imputing multiple times per tree on the subjects available in the respective node (see Table 2D).

TABLE 1 | Example table in a simulation setting. The columns with light gray background $\{\text{event}, T, C\}$ are produced by the data-generating mechanism but are not available during the training of the forest. The column *time* refers to $\tilde{T}_i = \min(T_i, C_i)$ and the column *status* is defined by $\Delta_i \cdot e_i$.

i	Time	Status	Event	T	C	X_1	X_2	X_3	...
1	15	1	1	15	18	−0.4411	−0.9011	−0.0924	...
2	7	2	2	7	12	−1.1834	0.7352	−0.1028	...
3	13	0	1	17	13	0.3930	−1.0282	1.2740	...
4	10	2	2	10	20	0.0181	−1.8797	−3.5290	...
5	3	1	1	3	14	0.7355	−1.0863	1.3222	...
...				...					

TABLE 2 | Illustration of data processing: Training time and status generated from data in Table 1. (A) The event time is replaced by the simulated (true) censoring time (usually not available for training in practice). (B) The competing event time is taken as censoring time, effectively ignoring the presence of competing events. (C) The censoring time ? is replaced once before fitting the forest by an estimated censoring time (based on weights w_{it} computed from the censoring survival function estimated from the entire training dataset). (D) The censoring time ?? is replaced repeatedly by an estimated censoring time based on weights w_{it} computed from the censoring survival function estimated from the training data subset available in the training data subset that is available in the specific node (root node) at the random forest.

	A: Simulated *C*			**B: Naive approach**			**C: Impute once**			**D: Impute in forest**	
i	Time	Status	i	Time	Status	i	Time	Status	i	Time	Status
1	15	1	1	15	1	1	15	1	1	15	1
2	12	0	2	7	0	2	?	0	2	??	0
3	20	0	3	13	0	3	13	0	3	13	0
4	10	0	4	10	0	4	?	0	4	??	0
5	3	1	5	3	1	5	3	1	5	3	1

In each simulation run, we divided the dataset into a training ($\frac{2}{3}$) and a test set ($\frac{1}{3}$) before applying the methods above. Splits were stratified by the event types (event of interest, competing event, censoring). We carried out 1000 simulation runs for each combination of setup, parameters q and b and for each imputation method, resulting in an overall number of 90,000 simulation runs. We chose 1000 runs because this number guaranteed the width of the reference limits for the CIF (provided in Figures S6 and S7) to be smaller than 0.1 (i.e., $2 \cdot 1.96 \cdot \sqrt{0.5 \cdot (1 - 0.5)/1000} = 0.0619 < 0.1$). Apart from the described incorporated imputation approaches, the RSFs were fitted using the R package **ranger** from the command line interface with default parameters. This means fitting $ntree = 500$ trees with $mtry = 8$ covariates selected in each node ($mtry = \sqrt{p}$), the log-rank split rule, sampling with replacement, and a minimal node size of 3.

2.5 | Illustration

To gain an understanding of the distribution of the imputed censoring times $\hat{C}_i$ in the imputeOnce method compared to the true censoring times C_i and the censoring times used in the naive approach ($\hat{C}_i = T_{e_i=2}$), Figure S3 depicts the distribution of C_i and $\hat{C}_i$ for one simulation run. As the censoring times are imputed multiple times in imputeNode and imputeRoot this visualization would be less meaningful, and we show the variation across life table estimates of G instead. To illustrate the variability of the life table estimates across trees and nodes, Figures S4 and S5 show examples for a setup, a simulation run, and a combination of b and q. Here we see that the estimation of G on the subsets in the trees (imputeRoot) leads to increased variability of $\hat{G}$. The estimation in each node of the trees (imputeNode) increases the variability even further. The mean squared error between the true censoring time C_i and the imputed censoring times $\hat{C}_i$ was lowest for imputeOnce and highest for imputeNode (see captions of Figures S3–S5).

2.6 | Performance Measures

2.6.1 | Calibration Graph

The agreement between the reference and estimated CIFs was evaluated using calibration graphs. Here, we directly compared the estimated CIF (averaged over the 1000 simulation runs) across the different (imputation) methods on the test dataset. The method containing the simulated true censoring times instead of the imputed censoring times served as a visual reference (see Table 2A). Generally, the methods are well calibrated if the averaged estimated CIF curve agrees closely with the reference.

2.6.2 | C-Index

The concordance index (C-index) was used to evaluate the discriminatory power of the different model fits on the test data. The

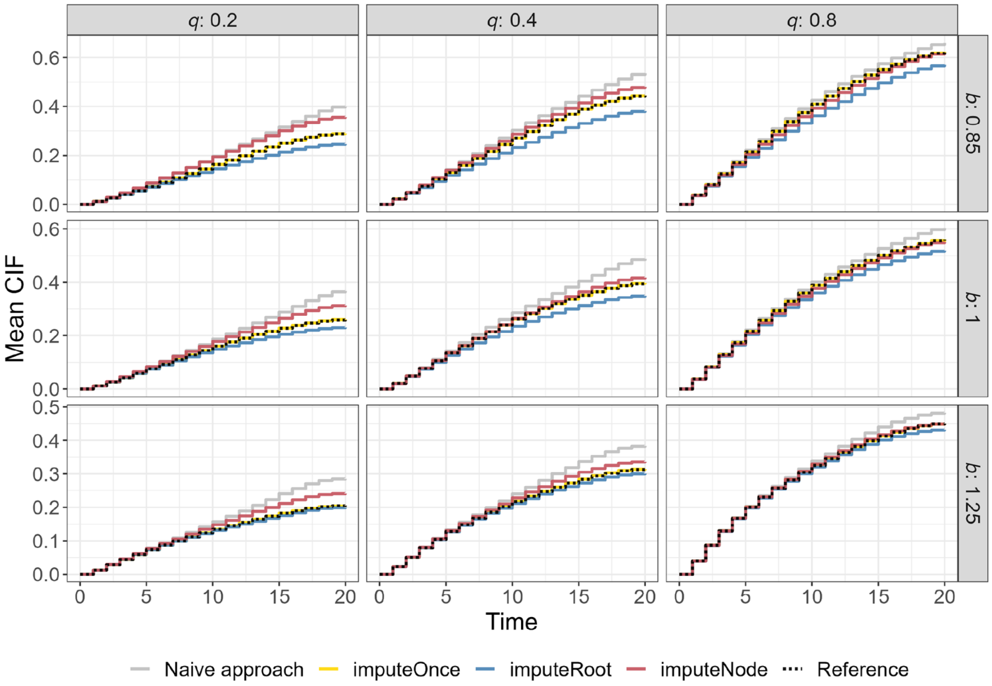

FIGURE 2 | Calibration graph for the test data of Setup 1 for different values of q (columns), determining the rate of the event of interest, and different censoring rates b (rows). A low value of q corresponds to a low rate of the event of interest, and a low value of b corresponds to a low censoring rate. In most scenarios, the black dotted line (reference based on true censoring times) visually overlaps with imputeOnce (yellow). (See Figure S1 for the relative frequencies of the event and censoring rates.) The CIF was averaged over 1000 simulation runs in each setting.

C-index essentially measures how well the ranking of the (time-averaged) estimated CHFs matches the ranking of the observed event times. A stronger alignment between these rankings with higher C-index values implies greater discriminatory power. The C-index as implemented in the function `cIndex` in the R package **discSurv** (Welchowski et al. 2022; Heyard et al. 2020) was calculated.

2.6.3 | Brier Score

The predictive performance of the approaches was compared using the Brier score (Gerds and Schumacher 2006). The Brier score at time point t is defined as the (estimated) squared difference between the observed and modeled status (Δ_i) at that time. The integrated Brier score (IBS) is calculated by integrating the Brier score over all possible time points t. Lower values imply a better prediction. The Brier score was calculated using the R package **pec** (Mogensen, Ishwaran, and Gerds 2012).

3 | Results

The calibration graphs in Figure 2 (simulation Setup 1) and Figure 3 (simulation setup 2) show the CIF on the test dataset that was not seen during training, averaged over 1000 simulation runs. They include nine different scenarios, that is, nine combinations of the parameters q and b, where q determines the rate of the event of interest, and b affects the censoring rate.

In all scenarios, all RSF architectures show similar CIF estimates for the first time points and tend to diverge for later time points. Here, the naive approach (gray lines), where competing events are treated as censoring, always shows the strongest overestimation and highest deviation from the reference method (dotted lines). The CIF of the imputeOnce approach visually overlaps with the dotted reference line that was obtained by training the single-event RSF on the simulated (true) censoring times (Reference).

The methods where the imputation is directly implemented in the nodes of the trees show the highest differences in the setting with a low censoring rate ($b = 0.85$, first row). In all settings, the method with only one imputation in each root node tends to underestimate the CIF. In contrast, the imputation in each node tends to overestimate the CIF, especially in the scenario that corresponds to a low event-of-interest rate and a low censoring rate ($q = 0.2, b = 0.85$). For a better understanding of the overlap of the estimated CIF, Figures S6 and S7 provide reference limits for the estimated CIF at time points 10, 15, and 20.

Concerning the C-index and the Brier score, all methods perform similarly Tables S1–S4). The methods that do not impute directly in the random forest (imputeOnce, naive approach) performed slightly better with regard to these metrics in Setup 2. However, in Setup 1, the imputation in the root nodes of the RSF (imputeRoot) performed similarly to imputeOnce. To further gain insight on the properties of the simulation design, we divided the 1000 simulation runs into 10 batches. Using these batches, an estimate

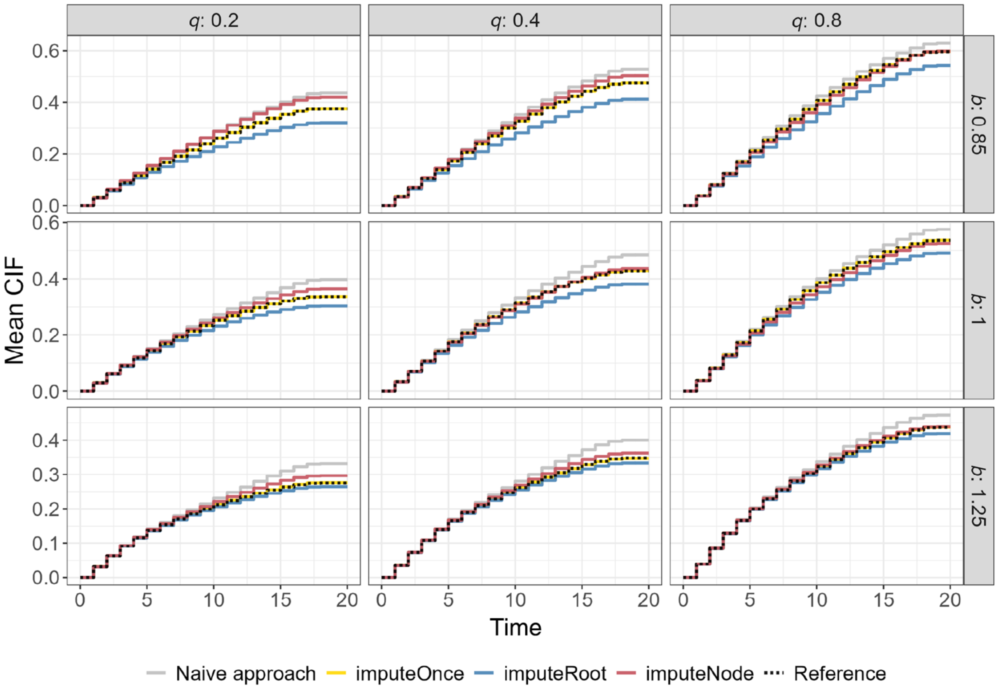

FIGURE 3 | Calibration graph for the test data of Setup 2 for different values of q (columns), determining to the rate of the event of interest, and different censoring rates b (rows). A low value of q corresponds to a low rate of the event of interest, and a low value of b corresponds to a low censoring rate. The CIF was averaged over 1000 simulation runs in each setting. In most scenarios, the black dotted line (reference with censoring times from simulation) visually overlaps with imputeOnce (yellow). (See Figure S2 for the relative frequencies of the event rates.)

of the Monte Carlo error was calculated. The corresponding results are presented in Figures S8–S15.

In addition to the described performance measures, we calculated the permutation variable importance (VIMP) on the training dataset using the time-aggregated CHF as a marker in Harrell's C-index (cf. Ishwaran et al. 2008). Figures S16–S20 show the 10 variables with the highest mean permutation VIMP averaged over 1000 simulation runs of the training datasets in Setup 1. Note that the covariates X_1 to X_5 were included in the data-generating mechanism for the event of interest (only the covariates X_4 and X_5 were associated on a continuous level), while the covariates X_1, X_3, X_4, X_6 were associated with the competing event (see Figure 1). The variables X_4 and X_5 are indeed the two most important variables throughout for the reference, the naive approach, and imputeOnce, while mostly only X_5 was considered in the first 10 variables for imputeRoot and imputeNode in the scenarios with a low and medium rate of the event of interest ($q \in \{0.2, 0.4\}$). In scenarios with a high rate of the event of interest ($q = 0.8$), X_2, X_3, X_4, X_5 were included for imputeRoot and imputeNode.

For Setup 2 (Figures S21–S25), the variables associated with the event of interest X_1 to X_5 are among the five most important variables in all scenarios for the Reference, the naive approach, and imputeOnce. In contrast, for the approaches imputeRoot and imputeNode, variables that are not associated with the event of interest get selected, especially in the scenarios with lower b. For imputeRoot and imputeNode, all of the variables X_1 to X_5 are only included in the first most important variables when the censoring rate is high ($b = 1.25$). For the high censoring scenario, the variables X_6 and X_7, which are associated with the competing event, are also in the top 10 most important variables.

3.1 | Limitations

We acknowledge that our simulation study has several limitations: First, our study did not have a preregistered study protocol. This was mainly because we designed our simulation study to gain insight into the properties of the proposed methodology and to provide the first empirical evidence on its functioning ("phase II" in the framework by Heinze et al. 2024). Clearly, more extended simulations covering a broader range of scenarios (corresponding to later phases in the framework by Heinze et al. 2024) will have to be based on preregistered protocols. Second, our simulation study used a rather limited set of values for the parameters k, q, and b. We chose these values because they had already been used in previous simulation studies with competing events (Beyersmann, Allignol, and Schumacher 2011; Berger et al. 2020), thus making our design consistent with earlier publications. Third, simulation Setups 1 and 2 were chosen to represent data-generating mechanisms with multiple interactions and arbitrary cut-offs, allowing us to mimic a scenario in which we typically would not fit a classical Cox proportional hazards model. These setups could be extended by data-generating mechanisms in which the competing event and the event of interest do not share risk factors (not explored in our simulations). They could further be extended to high-dimensional data settings where the number of covariates exceeds the number of observations.

4 | Application to the GCKD Study Data

We applied the methods described above to a subset of the GCKD study. In the observational, multicenter GCKD study (Titze et al. 2015), 5217 participants with CKD are followed up annually. Here, we look at data of up to 6.5 years of follow-up (data freeze: 03/2022), such that $k \in \{1, \dots, 7\}$, corresponding to 1-year intervals. We focus on one of the main events of interest in the GCKD study, namely reaching KF (dialysis, transplantation, or death due to forgoing kidney replacement therapy), while death by any other cause is considered a competing event (Table S5). More details on the data collection can be found in the Supporting Information (Section Application) and has been published, for example, in Steinbrenner et al. (2023). We included demographic and family history parameters as well as clinical and laboratory baseline parameters on categorical and continuous scales in the analyses. More specifically, we have considered the following baseline parameters:

- *Demographic*: age (in years), sex (male/female), alcohol (low-normal drinking/heavy drinking), smoking (nonsmokers/former smokers/smokers), family status (single/married or in a stable partnership/separated or divorced/widowed), number of siblings, number of people living in the household, employment (fully employed/part time/housework/pension/job-seeker/training/other), private insurance (yes/no), professional qualification (still in training/apprenticeship/master (craftsperson)/university degree/without degree/other/unknown);
- *Clinical*: enrollment (inclusion based on low eGFR value or proteinuria), body mass index (BMI, in kg/m^2), hypertension (yes/no), coronary heart disease (CHD: yes/no), stroke (yes/no), asthma (yes/no), chronic obstructive bronchitis (COPD: yes/no), taking painkillers (regularly/when required/never/unknown);
- *Laboratory*: serum creatinine (in mg/dL), eGFR (in mL/min $\cdot$ 1.73 m^2), UACR (in mg/g), CRP (in mg/L), low-density lipoprotein (LDL) cholesterol (in mg/dL), high-density lipoproetin (HDL) cholesterol (in mg/dL);
- *Family history*: number of siblings with stroke, number of siblings with kidney disease.

Further, diseases underlying CKD were dummy-coded for each participant (diabetic nephropathy, vascular nephropathy, systemic disease, primary glomerulopathy, interstitial nephropathy, acute kidney injury, single kidney, hereditary kidney disease, obstructive nephropathy, miscellaneous, undetermined). In many of the participants, more than one underlying disease was present, and a leading kidney disease was assigned by the treating nephrologist. Both the dummy encoded diseases underlying CKD and the assigned leading kidney disease are provided as covariates during the training of the forests, resulting in a total number of 38 covariates. Baseline characteristics are provided in Tables S6–S10. Note that several covariates are highly correlated, including individual and leading CKD causes and laboratory parameters. For example, the eGFR is calculated from the creatinine value, race, gender, and age using the CKD Epidemiology Collaboration (EPI) equation (Levey et al. 2009).

We compare the approaches described above on a complete case subset of the GCKD dataset. The dataset included 4256 participants. Of those, 412 (9.1%) reached KF (event of interest), and 409 (9.6%) died without reaching KF first (competing event, participants who died due to forgoing dialysis or transplantation are considered as KF). The estimated CIF and the 10 covariates with the highest VIMP can be seen in Figure 4. The CIF is lowest for the imputeOnce approach and imputeRoot. Due to the high sample size, imputeRoot and imputeOnce may lead to similar imputation results. We suspect the CIF of these two approaches to be the most realistic estimate based on the results of the simulation study, where the naive approach and imputeNode generally overestimated the CIF. Although the differences appear small, they will presumably become even more relevant with the longer observation period that can be evaluated in the future. The imputation method proposed by Ruan and Gray (2008) included analyses of multiple imputed datasets instead of a single imputation. Therefore, we performed 10 imputations of the imputeOnce method and compared the pooled results to single runs (see Table S11). In this application, however, the variability of the estimated CIF was quite low.

All approaches describe creatinine, UACR, eGFR, the leading CKD cause, and having a hereditary disease cause as the first five most important variables. This is followed by CRP, LDL cholesterol, and having diabetic nephropathy for imputeNode, imputeRoot, and naive approach. For imputeOnce, the demographic parameters age and sex were selected next instead of the laboratory parameters. The order of the selected covariates differs slightly between the approaches. A table containing VIMP values for all four methods can be found in Table S12. Both eGFR and UACR are reasonable covariates, as their progression is being discussed as a surrogate endpoint for progression to KF (Levey et al. 2020).

5 | Conclusion

We have proposed three variants of a subdistribution-based imputation approach to handle competing risks in RSF. Our simulation study showed that the CIF is well estimated when imputation already takes place outside the forest on the training data (imputeOnce).

In survival analysis, the occurrence of competing events must be appropriately taken into account. The naive approach of considering competing events as censoring can lead to biased estimates of the CIF, although our simulation study has shown that this approach may lead to similar results in terms of C-index and IBS. Differences in the estimated CIF became apparent, especially in scenarios with a high censoring rate or a low rate of the event of interest. By including the naive approach in the simulation, we wanted to raise awareness for the proper treatment of competing events when using machine learning applications.

It should be emphasized that the naive approach estimates the cause-specific CHF of the event of interest, ignoring the hazards of the competing events. Hence, it cannot be directly transformed

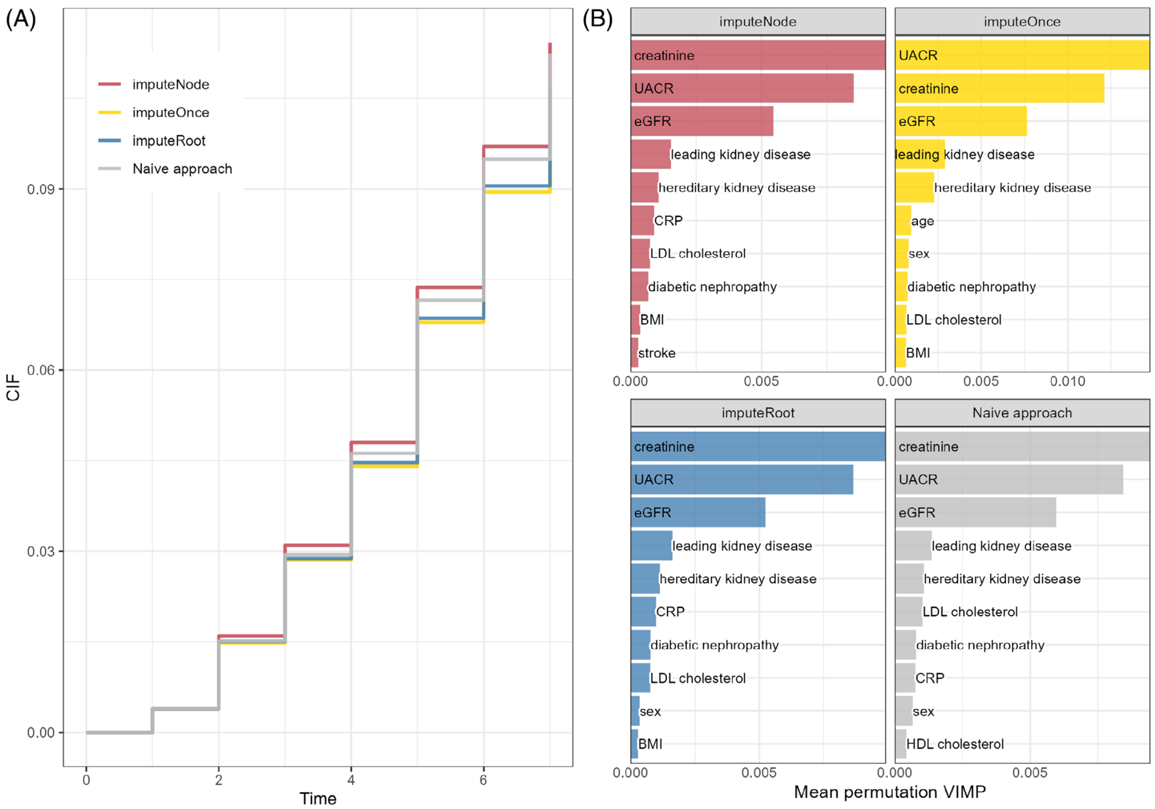

FIGURE 4 | (A) Estimated CIF when defining KF as the event of interest and death as the competing event on the GCKD dataset. Time is measured in years after baseline. Event times were discretized into 1-year intervals. (B) Mean permutation VIMP on the GCKD dataset for the different approaches. The 10 variables with the highest values are selected for each approach. The VIMP is calculated with respect to the prediction accuracy in the out-of-bag sample of the trees.

into the event of interest's CIF. While the CIF for the event of interest could be derived from a combination of all cause-specific hazard functions, we chose not to use this approach due to its complexity in analyzing covariate effects. Instead, we preferred the Fine and Gray method, as it provides a single (direct) effect per covariate. With random forests and other machine learning methods, having such a direct effect per covariate is a major advantage, in particular when it comes to the interpretation of measures like variable importance. Furthermore, the Fine and Gray method can reduce the computational effort, as it avoids having to fit separate machine learning models (one per cause-specific hazard). Also, note that the performance of the cause-specific hazard approach may strongly depend on the availability of sufficient numbers of observed events in the data.

A major finding of our simulation study is that imputing the estimated censoring times once before fitting the random forest (imputeOnce) essentially results in unbiased CIF estimates. Compared to imputations of the estimated censoring times in every tree node (imputeNode) or in the root node of the trees (imputeRoot), imputeOnce showed a systematically better performance with respect to the calibration graph of the CIF.

The question remains as to why the strategies imputeNode and imputeRoot resulted in an under/overestimation of the CIF in our simulation study. We considered the following two possible explanations:

i. In contrast to single imputation, with imputeNode, the sample sizes for estimating $G(t)$ are much smaller, especially in the direction of the terminal nodes, which are usually very small for RSF (default minimum node size: 3 in our simulation study). Therefore, the estimation of weights is less accurate, probably translating into less accurate, or even biased, estimates of CIF. In the imputeRoot scenario, the sample size is smaller than that of imputeOnce for subsamples, while with bootstrapping, there are additional problems due to ties, which can also lead to biases in the CIF estimates. We have seen this in the GCKD data: With a large sample size and a higher number of events, the differences between imputeOnce and imputeRoot are smaller and the estimate of $G(t)$ stabilizes.

ii. With imputeNode, the censoring survival function $G(t)$ is reestimated in each node and thus on the subset of data that is available in the specific node. Consequently, due to smaller sample sizes in the lower levels of the trees, imputeNode tends to show much higher variability in the estimation of the censoring survival function than imputeOnce. The censoring times might thus be imputed with reduced precision, resulting in a decreased estimation accuracy of the

CIF. Similar arguments hold for the imputeRoot strategy (effectively operating on data samples with a reduced size).

In conclusion, the proposed single-imputation strategy (imputeOnce) allows for converting the competing-risks setting into a single-event setting. All RSF features and options (split rules, variable importance measure, etc.) are immediately available for this setting, making it much more straightforward to apply RSF in the competing-risks context. Issues for future research include a comparison to other machine learning methods and other techniques for dealing with competing events in RSF. This could, for example, be done in the framework of a neutral comparison study (see, e.g., the recently published Special Collection on "Neutral Comparison Studies in Methodological Research" in Vol. 66 of *Biometrical Journal*).

Acknowledgments

M.N.W. was supported by the German Research Foundation (DFG) - Emmy Noether Grant 437611051. The German Chronic Kidney Disease (GCKD) study was funded by grants from the German Federal Ministry of Education and Research (BMBF, grant number 01ER0804), the KfH Foundation for Preventive Medicine, and corporate sponsors (http://www.gckd.org). We are grateful for the willingness of the participants to participate in the GCKD study. The enormous effort of the study personnel of the various regional centers is highly appreciated. We thank the large number of nephrologists who provide routine care for the participants and collaborate with the GCKD study. A list of nephrologists currently collaborating with the GCKD study is available at http://www.gckd.org.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The code for modifying the RSF implementation is available at https://github.com/cbehning/ranger/tree/competing_risks_subdist. The code for replicating the simulation results is available at https://github.com/cbehning/rsf_competing_events

GCKD: Public posting of individual-level participant data is not covered by the informed patient consent form. As stated in the patient consent form and approved by the Ethics Committees, a dataset containing pseudonyms can be obtained by collaborating scientists upon approval of a scientific project proposal by the steering committee of the GCKD study: https://www.gckd.org.

Open Research Badges

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to data confidentiality issues.

References

Archer, K. J., and R. V. Kimes. 2008. "Empirical Characterization of Random Forest Variable Importance Measures." *Computational Statistics & Data Analysis* 52, no. 4: 2249–2260.

Beck, H., S. I. Titze, S. Hübner, et al. 2015. "Heart Failure in a Cohort of Patients With Chronic Kidney Disease: The GCKD Study." *PLoS ONE* 10, no. 4: e0122552.

Berger, M., M. Schmid, T. Welchowski, S. Schmitz-Valckenberg, and J. Beyersmann. 2020. "Subdistribution Hazard Models for Competing Risks in Discrete Time." *Biostatistics* 21, no. 3: 449–466.

Beyersmann, J., A. Allignol, and M. Schumacher. 2011. *Competing Risks and Multistate Models With R*. New York: Springer Science & Business Media.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45: 5–32.

Cox, D. R. 1972. "Regression Models and Life-Tables." *Journal of the Royal Statistical Society: Series B (Methodological)* 34, no. 2: 187–202.

Fine, J. P., and R. J. Gray. 1999. "A Proportional Hazards Model for the Subdistribution of a Competing Risk." *Journal of the American Statistical Association* 94, no. 446: 496–509.

Gerds, T. A., and M. Schumacher. 2006. "Consistent Estimation of the Expected Brier Score in General Survival Models With Right-Censored Event Times." *Biometrical Journal* 48: 1029–1040.

Giunchiglia, E., A. Nemchenko, and M. van der Schaar. 2018. "RNN-SURV: A Deep Recurrent Model for Survival Analysis." In *Proceedings of the 27th International Conference on Artificial Neural Networks*, 23–32. Cham: Springer.

Gorgi Zadeh, S., C. Behning, and M. Schmid. 2022. "An Imputation Approach Using Subdistribution Weights for Deep Survival Analysis With Competing Events." *Scientific Reports* 12, no. 1: 3815.

Gupta, G., V. Sunder, R. Prasad, and G. Shroff. 2019. "Cresa: A Deep Learning Approach to Competing Risks, Recurrent Event Survival Analysis." In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 108–122. Cham: Springer.

Heinze, G., A.-L. Boulesteix, M. Kammer, T. P. Morris, I. R. White, and the Simulation Panel of the STRATOS Initiative. 2024. "Phases of Methodological Research in Biostatistics–Building the Evidence Base for New Methods." *Biometrical Journal* 66, no. 1: 2200222.

Heyard, R., J.-F. Timsit, L. Held, and COMBACTE-MAGNET Consortium. 2020. "Validation of Discrete Time-to-Event Prediction Models in the Presence of Competing Risks." *Biometrical Journal* 62, no. 3: 643–657.

Hothorn, T., B. Lausen, A. Benner, and M. Radespiel-Tröger. 2004. "Bagging Survival Trees." *Statistics in Medicine* 23, no. 1: 77–91.

Hsu, J. Y., J. A. Roy, D. Xie, et al. 2017. "Statistical Methods for Cohort Studies of CKD: Survival Analysis in the Setting of Competing Risks." *Clinical Journal of the American Society of Nephrology* 12, no. 7: 1181–1189.

Ishwaran, H., T. A. Gerds, U. B. Kogalur, R. D. Moore, S. J. Gange, and B. M. Lau. 2014. "Random Survival Forests for Competing Risks." *Biostatistics* 15, no. 4: 757–773.

Ishwaran, H., U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. 2008. "Random Survival Forests." *The Annals of Applied Statistics* 2: 841–860.

Lee, C., W. R. Zame, J. Yoon, and M. van der Schaar. 2018. "DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks." In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2314–2321. Palo Alto, CA: AAAI Press.

Levey, A. S., R. T. Gansevoort, J. Coresh, et al. 2020. "Change in Albuminuria and GFR as End Points for Clinical Trials in Early Stages of CKD: A Scientific Workshop Sponsored by the National Kidney Foundation in Collaboration With the US Food and Drug Administration and European Medicines Agency." *American Journal of Kidney Diseases* 75, no. 1: 84–104.

Levey, A. S., L. A. Stevens, C. H. Schmid, et al. 2009. "A New Equation to Estimate Glomerular Filtration Rate." *Annals of Internal Medicine* 150, no. 9: 604–612.

Mogensen, U. B., and T. A. Gerds. 2013. "A Random Forest Approach for Competing Risks Based on Pseudo-Values." *Statistics in Medicine* 32, no. 18: 3102–3114.

Mogensen, U. B., H. Ishwaran, and T. A. Gerds. 2012. “Evaluating Random Forests for Survival Analysis Using Prediction Error Curves.” *Journal of Statistical Software* 50, no. 11: 1–23.

Ren, K., J. Qin, L. Zheng, et al. 2019. “Deep Recurrent Survival Analysis.” In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 4798–4805. Palo Alto, CA: AAAI Press.

Ruan, P. K., and R. J. Gray. 2008. “Analyses of Cumulative Incidence Functions via Non-Parametric Multiple Imputation.” *Statistics in Medicine* 27, no. 27: 5709–5724.

Schmid, M., T. Welchowski, M. N. Wright, and M. Berger. 2020. “Discrete-Time Survival Forests With Hellinger Distance Decision Trees.” *Data Mining and Knowledge Discovery* 34, no. 3: 812–832.

Schmid, M., M. N. Wright, and A. Ziegler. 2016. “On the Use of Harrell’s C for Clinical Risk Prediction via Random Survival Forests.” *Expert Systems With Applications* 63: 450–459.

Steinbrenner, I., P. Sekula, F. Kotsis, et al. 2023. “Association of Osteopontin With Kidney Function and Kidney Failure in Chronic Kidney Disease Patients: The GCKD Study.” *Nephrology Dialysis Transplantation* 38, no. 6: 1430–1438.

Therrien, J., and J. Cao. 2022. “Random Competing Risks Forests for Large Data.” arXiv preprint arXiv:2207.11590.

Titze, S., M. Schmid, A. Köttgen, et al. 2015. “Disease Burden and Risk Profile in Referred Patients With Moderate Chronic Kidney Disease: Composition of the German Chronic Kidney Disease (GCKD) Cohort.” *Nephrology Dialysis Transplantation* 30, no. 3: 441–451.

Welchowski, T., M. Berger, D. Koehler, and M. Schmid. 2022. *discSurv: Discrete Time Survival Analysis*. R Package Version 2.0.0. https://CRAN.R-project.org/package=discSurv.

Wright, M. N., T. Dankowski, and A. Ziegler. 2017. “Unbiased Split Variable Selection for Random Survival Forests Using Maximally Selected Rank Statistics.” *Statistics in Medicine* 36, no. 8: 1272–1284.

Wright, M. N., and A. Ziegler. 2017. “Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R.” *Journal of Statistical Software* 77, no. 1: 1–17. https://doi.org/10.18637/jss.v077.i01.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

4 Discussion

The articles in this dissertation address aspects of planning and analyzing longitudinal data using novel approaches for statistical learning and regression. Due to the work on the MACUSTAR study during the dissertation period, the statistical methods presented here cover topics relevant to longitudinal research on age-related macular degeneration. First, factors influencing the recruitment of participants into a longitudinal observational study were evaluated using a Poisson regression model. Second, a novel regression framework for disease progression and unknown disease onset was presented. The third and fourth articles presented imputation approaches for the evaluation of time-to-event endpoints in the presence of competing events in statistical learning methods.

4.1 Understanding early disease progression

Recruitment failure can result in reduced power and has ethical, scientific, and financial implications (McDonald et al., 2006; Sully et al., 2013). The findings from the MACUSTAR study as presented in Terheyden et al. (2021) showed that increased teleconferencing with site investigators, public holidays, and reaching 80% of impaired screening performance impacted recruitment rates. These factors should be carefully considered in future study designs and site selection, especially when recruiting early, asymptomatic diseased participants.

The mixed-model framework presented in Behning et al. (2021) found that a square-root transformation is a reasonable choice to model enlargement of GA lesions. While several studies in manifested GA, including interventional phase 2 and 3 trials, have been using the square-root-transformation to study progression (Steinle et al., 2021; Khanani et al., 2023; Keenan et al., 2024a; Keenan et al., 2024b), the trajectory and associated risk factors of early GA smaller than the often used minimum lesion size requirement for clinical trials (e.g., 0.5 mm^2) is poorly understood. Further research is needed to study the preceding AMD disease states, such as incomplete to complete retinal pigment epithelium and outer retinal atrophy (iRORA and cRORA), and nascent GA (Wu et al., 2020; Rajanala et al., 2023).

Based on the proposed modeling framework in Behning et al. (2021), the (unknown) age at GA onset was estimated using the information from baseline covariates. In the future, additional risk factors could be included in order to identify the age of GA onset more precisely. This can help to assess the influence of genetic and environmental factors and design more targeted clinical trials that consider the specific needs of different patient subgroups.

To summarize, the findings derived from Publications A and B help to define study inclusion criteria and facilitate screening for future clinical trials in multi-center settings for disease stages preceding GA. In addition, the modeling framework in Publication B allows for adequate modeling of continuous endpoints in study populations with unknown age of onset and unknown progression patterns (e.g., linear, quadratic, exponential).

4.2 Time-to-event endpoints and competing risks

To date, no clinical trial endpoints have been validated and accepted as clinical endpoints by regulatory agencies for drug development in early AMD-stages (Finger et al., 2019). At present, both continuous and time-to-event endpoints are potentially suitable for future clinical trials. Should future clinical trials in intermediate AMD populations employ time-to-event endpoints, possible occurrences of competing events must be carefully considered.

Publications C and D included in this dissertation demonstrated the importance of analyzing competing events in longitudinal data and showed a practical option for addressing them in statistical learning methods.

The imputation strategy proposed by Ruan and Gray (2008) bears some similarities to the imputation strategy in Publications C and D. However, their method was developed in a continuous-time framework, used multiple imputations, and applied to a Cox proportional hazards model. Here, we applied a related imputation method to statistical learning methods. As many implementations of statistical learning methods treat time as an ordinal variable and thus use discrete-time data structures (Lee et al., 2018; Ren et al., 2018; Ishwaran et al., 2008), both publications considered a discrete-time modeling framework. We showed that the imputation strategy using subdistribution weights transformed competing event survival data so that it can be used in single-event statistical learning methods. We have demonstrated this approach both in deep survival network architectures (Gorgi Zadeh et al., 2022) and single-event random survival forests (Behning et al., 2024a).

The imputation of unknown censoring times in a preprocessing step provides a practical solution to avoid biased estimation results and predictions in the presence of competing events. Well-established ML architectures for single events can easily be applied to competing event data that may not have been initially considered during the study planning. Training a single subdistribution survival ML model facilitates the fitting and interpretation of associated risk factors compared to training multiple cause-specific architectures, especially as the performance of cause-specific architectures may also depend on the availability of sufficiently large numbers of observations per event type.

Publication C (Gorgi Zadeh et al., 2022) and Publication D (Behning et al., 2024a) provide application examples using data from longitudinal medical studies, more specifically for data sets from oncology, emergency medicine, and nephrology. Although these methods were not applied in the field of ophthalmology, they can also be utilized in future analysis of longitudinal studies in ophthalmology. For example, these methods can be applied to identify risk factors associated with faster progression to late-stage AMD as collected in the MACUSTAR study. Progression was modeled in discrete time based on the six-monthly visits in the study (Finger et al., 2019; Behning et al., 2024b; Dunbar et al., 2024; Sassmannshausen et al., 2024). The DNN approach can be trained with semi-structured baseline data, such as images arising from optical coherence tomography. The RSF approach can be beneficial for structured high-dimensional data, e.g., combining multiple structural, functional, genetic, or patient-reported outcome measures at baseline.

While this dissertation implemented new imputation methods for DNN and RSF, an extension to other statistical learning methods could also be possible. In fact, the imputation step can already be carried out as part of the data preprocessing and is largely independent of the subsequently applied statistical learning method. The imputation can also offer a practical alternative if no competing risk implementations are yet available for a particular ML method.

4.3 Conclusion

In summary, the four publications of this dissertation contribute to improving the planning and analysis of longitudinal studies, with a focus on application in AMD research. As no regulatory accepted endpoint exists for clinical trials in earlier stages of AMD, it is necessary to investigate the statistical methods used for both longitudinal change endpoints and time-to-event endpoints. Crucially, if future studies consider GA and CNV, the two late stages of AMD, as separate time-to-event endpoints, an awareness of the correct treatment of competing risks is important. While this dissertation focused on ophthalmological research, the presented methods are not limited to this field but can also be applied to other slowly progressing diseases, e.g., chronic kidney disease.

4.4 References

Behning C, Bigerl A, Wright MN, Sekula P, Berger M, Schmid M. Random Survival Forests With Competing Events: A Subdistribution-Based Imputation Approach. In: Biometrical Journal 2024; 66 (6): e202400014

Behning C, Fleckenstein M, Pfau M, Adrion C, Goerdt L, Lindner M, Schmitz-Valckenberg S, Holz FG, Schmid M. Modeling of atrophy size trajectories: variable transformation, prediction and age-of-onset estimation. In: BMC Medical Research Methodology 2021; 21: 1–12

Behning C, Terheyden JH, Finger R, Dunbar HM, Sassmannshausen M, Tufail A, Crabb DP, Binns AM, Wu Z, Guymer RH, et al. Validating a Confirmatory Structure-Function Model Prognostic of Progression of Intermediate AMD: MACUSTAR Study Primary Endpoint Analysis. In: Investigative Ophthalmology & Visual Science 2024; 65 (7): 4314–4314

Dunbar HM, Behning C, Binns AM, Terheyden JH, Poor SH, Finger R, Crabb DP, Leal S, Tufail A, Holz FG, et al. The prognostic power of baseline visual function deficits in intermediate age-related macular degeneration (iAMD) for progression to late AMD-A MACUSTAR study report. In: Investigative Ophthalmology & Visual Science 2024; 65 (7): 1485–1485

Finger RP, Schmitz-Valckenberg S, Schmid M, Rubin GS, Dunbar H, Tufail A, Crabb DP, Binns A, Sánchez CI, Margaron P, et al. MACUSTAR: development and clinical validation of functional, structural, and patient-reported endpoints in intermediate age-related macular degeneration. In: Ophthalmologica 2019; 241 (2): 61–72

Gorgi Zadeh S, Behning C, Schmid M. An imputation approach using subdistribution weights for deep survival analysis with competing events. In: Scientific Reports 2022; 12 (1): 3815

Ishwaran H, Kogalur U, Blackstone EH, Lauer MS. Random survival forests. In: The Annals of Applied Statistics 2008; 2: 841–860

Keenan TD, Agrón E, Keane PA, Domalpally A, Chew EY, et al. Oral Antioxidant and Lutein/Zeaxanthin Supplements Slow Geographic Atrophy Progression to the Fovea in Age-Related Macular Degeneration. In: Ophthalmology 2024:

Keenan TD, Bailey C, Abraham M, Orndahl C, Menezes S, Bellur S, Arunachalam T, Kangale-Whitney C, Srinivas S, Karamat A, et al. Phase 2 Trial Evaluating Minocycline for Geographic Atrophy in Age-Related Macular Degeneration: A Nonrandomized Controlled Trial. In: JAMA ophthalmology 2024; 142 (4): 345–355

Khanani AM, Patel SS, Staurenghi G, Tadayoni R, Danzig CJ, Eichenbaum DA, Hsu J, Wykoff CC, Heier JS, Lally DR, et al. Efficacy and safety of avacincaptad pegol in patients with geographic atrophy (GATHER2): 12-month results from a randomised, double-masked, phase 3 trial. In: The Lancet 2023; 402 (10411): 1449–1458

Lee C, Zame WR, Yoon J, Schaar M van der. DeepHit: A deep learning approach to survival analysis with competing risks. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018: 2314–2321

McDonald AM, Knight RC, Campbell MK, Entwistle VA, Grant AM, Cook JA, Elbourne DR, Francis D, Garcia J, Roberts I, et al. What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. In: Trials 2006; 7: 1–8

Rajanala K, Dotiwala F, Upadhyay A. Geographic atrophy: Pathophysiology and current therapeutic strategies. In: Frontiers in Ophthalmology 2023; 3: 1327883

Ren K, Qin J, Zheng L, Yang Z, Zhang W, Qiu L, Yu Y. Deep recurrent survival analysis. Tech. rep. Technical report, arXiv:1809.02403v2 [cs.LG]

Ruan PK, Gray RJ. Analyses of cumulative incidence functions via non-parametric multiple imputation. In: Statistics in Medicine 2008; 27 (27): 5709–5724

Sassmannshausen M, Thiele S, Terheyden JH, Luhmann UF, Leal S, Schmid M, Finger R, Holz FG, Schmitz-Valckenberg S, Behning C. Prognostic value of structural biomarkers for disease progression in intermediate age-related macular degeneration: a MACUSTAR study report. In: Investigative Ophthalmology & Visual Science 2024; 65 (7): 5684–5684

Steinle NC, Pearce I, Monés J, Metlapally R, Saroj N, Hamdani M, Ribeiro R, Rosenfeld PJ, Lad EM. Impact of baseline characteristics on geographic atrophy progression in the FILLY trial evaluating the complement C3 inhibitor pegcetacoplan. In: American journal of ophthalmology 2021; 227: 116–124

Sully BG, Julious SA, Nicholl J. A reinvestigation of recruitment to randomised, controlled, multicenter trials: a review of trials funded by two UK funding agencies. In: Trials 2013; 14: 1–9

Terheyden JH, Behning C, Lüning A, Wintergerst L, Basile PG, Tavares D, Melício BA, Leal S, Weissgerber G, Luhmann UF, et al. Challenges, facilitators and barriers to screening study participants in early disease stages-experience from the MACUSTAR study. In: BMC Medical Research Methodology 2021; 21: 1–8

Wu Z, Luu CD, Hodgson LA, Caruso E, Tindill N, Aung KZ, McGuinness MB, Makeyeva G, Chen FK, Chakravarthy U, et al. Prospective longitudinal evaluation of nascent geographic atrophy in age-related macular degeneration. In: Ophthalmology Retina 2020; 4 (6): 568–575

Additional research references

MACUSTAR

Dunbar HM, Behning C, Abdirahman A, Higgins BE, Binns AM, Terheyden JH, Zakaria N, Poor S, Finger RP, Leal S, et al. Repeatability and discriminatory power of chart-based visual function tests in individuals with age-related macular degeneration: a MACUSTAR study report. In: JAMA ophthalmology 2022; 140 (8): 780–789

Higgins BE, Montesano G, Dunbar HM, Binns AM, Taylor DJ, Behning C, Abdirahman A, Schmid MC, Terheyden JH, Zakaria N, et al. Test-retest variability and discriminatory power of measurements from microperimetry and dark adaptation assessment in people with intermediate age-related macular degeneration–a MACUSTAR study report. In: Translational Vision Science & Technology 2023; 12 (7): 19–19

Saßmannshausen M, Behning C, Isselmann B, Schmid M, Finger RP, Holz FG, Schmitz-Valckenberg S, Pfau M, Thiele S. Relative ellipsoid zone reflectivity and its association with disease severity in age-related macular degeneration: a MACUSTAR study report. In: Scientific reports 2022; 12 (1): 14933

Saßmannshausen M, Behning C, Weinz J, Goerdt L, Terheyden JH, Chang P, Schmid M, Poor SH, Zakaria N, Finger RP, et al. Characteristics and spatial distribution of structural features in age-related macular degeneration: a MACUSTAR study report. In: Ophthalmology Retina 2023; 7 (5): 420–430

Terheyden JH, Schmitz-Valckenberg S, Crabb DP, Dunbar H, Luhmann UFO, Behning C, Schmid M, Silva R, Cunha-Vaz J, Tufail A, et al. Use of composite end points in early and intermediate age-related macular degeneration clinical trials: state-of-the-art and future directions. In: Ophthalmologica 2021; 244 (5): 387–395

Terheyden JH, Pondorfer SG, Behning C, Berger M, Carlton J, Rowen D, Bouchet C, Poor S, Luhmann UF, Leal S, et al. Disease-specific assessment of Vision Impairment in Low Luminance in age-related macular degeneration–a MACUSTAR study report. In: British Journal of Ophthalmology 2023; 107 (8): 1144–1150

Other projects

Bauer CJ, Karakostas P, Weber N, Behning C, Stoffel-Wagner B, Brossart P, Dolscheid-Pommerich R, Schäfer VS. Comparative analysis of contemporary anti-double stranded DNA antibody assays for systemic lupus erythematosus. In: Frontiers in Immunology 2023; 14: 1305865

Bigerl A, Conrads F, Behning C, Sherif MA, Saleem M, Ngonga Ngomo AC. Tentris–a tensor-based triple store. In: Pan J et al., eds. International Semantic Web Conference, Springer, 2020: 56–73

Bigerl A, Conrads L, Behning C, Saleem M, Ngonga Ngomo AC. Hashing the hypertrie: space- and time-efficient indexing for SPARQL in tensors. In: Sattler U et al., eds. International Semantic Web Conference, Springer, 2022: 57–73

Burg LC, Karakostas P, Behning C, Brossart P, Kermani TA, Schäfer VS. Prevalence and characteristics of giant cell arteritis in patients with newly diagnosed polymyalgia rheumatica–a prospective cohort study. In: Therapeutic Advances in Musculoskeletal Disease 2023; 15: 1759720X221149963

Grobelski J, Wilsmann-Theis D, Karakostas P, Behning C, Brossart P, Schäfer VS. Prospective double-blind study on the value of musculoskeletal ultrasound by dermatologists as a screening instrument for psoriatic arthritis. In: Rheumatology 2023; 62 (8): 2724–2731

Karakostas P, Dejaco C, Behning C, Recker F, Schäfer VS. Point-of-care ultrasound enables diagnosis of giant cell arteritis with a modern innovative handheld probe. In: Rheumatology 2021; 60 (9): 4434–4436

Kravchenko D, Behning C, Bergner R, Schäfer VS. How to Differentiate Gout, Calcium Pyrophosphate Deposition Disease, and Osteoarthritis Using Just Four Clinical Parameters. In: Diagnostics 2021; 11 (6): 924

Kravchenko D, Karakostas P, Kuetting D, Meyer C, Brossart P, Behning C, Schäfer VS. The role of dual energy computed tomography in the differentiation of acute gout flares and acute calcium pyrophosphate crystal arthritis. In: Clinical rheumatology 2022: 1–11

Langner SM, Terheyden JH, Geerling CF, Kindler C, Keil VC, Turski CA, Turski GN, Behning C, Wintergerst MW, Petzold GC, et al. Structural retinal changes in cerebral small vessel disease. In: Scientific reports 2022; 12 (1): 9315

Mockenhaupt LM, Dolscheid-Pommerich R, Stoffel-Wagner B, Behning C, Brossart P, Schäfer VS. Autoantibodies to dense-fine-speckled 70 (DFS70) do not necessarily rule out connective tissue diseases. In: Seminars in Arthritis and Rheumatism 2022; 52: 151936

Petzinna SM, Burg LC, Bauer CJ, Karakostas P, Terheyden JH, Behning C, Holz FG, Brossart P, Finger RP, Schäfer VS. Transorbital ultrasound in the diagnosis of giant cell arteritis. In: Rheumatology 2024: keae287

Petzinna SM, Winter L, Skowasch D, Pizarro C, Weber M, Kütting D, Behning C, Bauer CJ, Schäfer VS. Assessing sleep-related breathing disorders among newly diagnosed rheumatoid and psoriatic arthritis patients: a cross-sectional study. In: Rheumatology International 2024; 44 (6): 1025–1034

Plöger R, Behning C, Walter A, Jimenez Cruz J, Gembruch U, Strizek B, Recker F. Next-generation monitoring in obstetrics: Assessing the accuracy of non-piezo portable ultrasound technology. In: Acta Obstetricia et Gynecologica Scandinavica 2024:

Schäfer VS, Dejaco C, Karakostas P, Behning C, Brossart P, Burg LC. Follow-up ultrasound examination in patients with newly diagnosed giant cell arteritis. In: Rheumatology 2024: keae098

Schreiner JK, Recker F, Scheicht D, Karakostas P, Ziob J, Behning C, Preuss P, Brossart P, Schäfer VS. Changes in ultrasound imaging of joints, entheses, bursae and tendons 24 and 48 h after adjusted weight training. In: Therapeutic Advances in Musculoskeletal Disease 2022; 14: 1759720X221111610

Terheyden JH, Ost RA, Behning C, Mekschrat L, Bildik G, Wintergerst MW, Holz FG, Finger RP. Evaluation of the test–retest and inter-mode comparability of the Impact of Vision Impairment questionnaire in people with chronic eye diseases. In: Graefe's Archive for Clinical and Experimental Ophthalmology 2024; 262 (6): 1933–1943

Verspohl SH, Holderried T, Behning C, Brossart P, Schäfer VS. Prevalence, therapy and tumour response in patients with rheumatic immune-related adverse events following immune checkpoint inhibitor therapy: a single-centre analysis. In: Therapeutic Advances in Musculoskeletal Disease 2021; 13: 1759720X211006963

Vychopen M, Hamed M, Bahna M, Racz A, Ilic I, Salemdawod A, Schneider M, Lehmann F, Eichhorn L, Bode C, et al. A Validation Study for SHE Score for Acute Subdural Hematoma in the Elderly. In: Brain Sciences 2022; 12 (8): 981

Vychopen M, Schneider M, Borger V, Schuss P, Behning C, Vatter H, Güresir E. Complete hemispheric exposure vs. superior sagittal sinus sparing craniectomy: incidence of shear-bleeding and shunt-dependency. In: European Journal of Trauma and Emergency Surgery 2022: 1–9

Ziob J, Behning C, Brossart P, Bieber T, Wilsmann-Theis D, Schäfer VS. Specialized dermatological-rheumatological patient management improves diagnostic outcome and patient journey in psoriasis and psoriatic arthritis: a four-year analysis. In: BMC rheumatology 2021; 5: 1–8

5 Acknowledgements

First, I would like to thank Prof. Dr. Matthias Schmid for his valuable feedback and productive ideas on my work. Further, I greatly appreciate the support I received from the cooperation with the Institute of Ophthalmology and everyone working on the MACUSTAR study. My sincere gratitude goes to the members of my dissertation committee, Prof. Dr. Inke König, Prof. Dr. Andreas Mayr and Prof. Dr. Robert Finger. I want to thank all of my former and present colleagues at IMBIE for constructive discussions, entertaining lunches, and the enjoyable working atmosphere. In particular, I would like to thank Leonie, Marie, Dario, and Hannah for their constructive feedback when helping me out at short notice as proofreaders of my thesis. Thanks to Jan, I learned a lot about the formalities of the doctoral office from you. Special thanks go to Moritz and Shekoufeh for all the helpful discussions, words of encouragement, and fun talks that we had in our office room.

Many thanks also to my colleagues from ophthalmology, especially to Hannah and Jan, for their open ears, the many great discussions, valuable feedback on this thesis, and the fun we had at conferences.

A heartfelt thanks to my friends and family for their unconditional support and confidence in me and for reminding me that there are things more important than work. Thanks to Alex for your patience, your encouragement, your support, for everything.