# Overview and comparison of methods

## Introduction

Let $T$ denote the time to event, and $\Delta \in \{1, \dots, d\}$ denote the indicator of the cause of the event for $d$ competing causes. In analyses of these sort of data, it may be of interest to estimate and model the cause specific cumulative incidence of cause 1 (WOLOG) at a particular time $t^*$:

$$P(T < t^*, \Delta = 1)$$

or the restricted mean life years lost due to cause 1 up to time $t^*$:

$$\int_0^{t^*} P(T < u, \Delta = 1)\, du.$$

Andersen et al. [2013]

## Methods of calculating pseudo-observations

### The OG pseudo-observation approach

Andersen et al. [2003] developed the original approach using the jackknife. Let $\theta$ denote the parameter of interest and $\hat{\theta}$ the estimate using all of the observations. Let $\hat{\theta}_{-i}$ denote the estimate obtained by leaving the $i$th observation out of the sample and recomputing the estimate. Then the $i$th pseudo-observation is $P_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}$.

When $\theta$ is the cumulative incidence and the estimate is based on the Aalen-Johansen estimator, then there are some computational tricks so that the estimator does not need to be rerun $n$ times. This approach is implemented in the prodlim package [Gerds, 2019] with some slight modifications by me to be more memory saving when there is a large dataset. In the case of the restricted mean, no such tricks are readily implemented and we recompute the Aalen-Johansen $n$ times and integrate each time.

Andersen and Pohar Perme [2010]

### The infintesimal jackknife approach

We can rewrite the $i$th pseudo-observation as $P_i = \hat{\theta} + (n-1)(\hat{\theta} - \hat{\theta}_{-i})$, and note that $(\hat{\theta} - \hat{\theta}_{-i}) \approx \partial\hat{\theta}/\partial w_i$, where $w_i$ is the case weight for subject $i$. The right side of that equation is approximated by a Taylor series expansion, and is in fact the $i$ subjects contribution to the empirical influence function of $\theta$. Calculation of the influence function contributions is done already by the survival package [Therneau, 2020], and returned to the user as of version 3.0. We use these influence functions in the calculation of $P_i$ as above for the cumulative incidence. The calculation for the restricted mean is slightly more involved because it involves all influence functions up to time $t^*$, but no additional recalculation is needed.

Expand on this because it is being described for the first time Jaeckel [1972], Efron [1992]

### Inverse probability of censoring weighted

We have also implemented the pseudo observation estimators as described by Overgaard et al. [2019]. We allow the option of either a Cox model of Aalen's additive hazards model for estimating the probabilities of remaining uncensored.

## Regression models

### Advantages over Cox regression

Collapsibility of the risk difference/risk ratio. Causal inference.

Interpretability in terms of absolute risk.

No need to assume proportional hazards.

### Estimation

### Variance estimation

## Simulation study

We conducted a simulation study with the goal of determining which methods should be used as the defaults in our package. The key criteria are validity, as measured by type I error rates, bias, and confidence interval coverage, robustness to misspecification of the censoring mechanism, and statistical efficiency. A lesser concern is computational efficiency.

We generated datasets with competing risks according to Beyersmann et al. [2009] as follows: We first generated a binary covariate $Z$ as Bernoulli with probability 0.5, and two independent standard normal variables $X_1, X_2$. Then $\mathbf{Q} = (1, Z, X_1, X_2)$. We used a proportional hazards Weibull distribution to generate the time data for $k = 1, 2$, with a hazard of: $h_k(t|\mathbf{Q}) = \gamma_k * (1/e^{(\mathbf{Q}^T\zeta_k)})^{\gamma_k} * t^{\gamma_k - 1}$ and a cumulative hazard given by: $H_k(t|\mathbf{Q}) = (1/e^{(\mathbf{Q}^T\zeta_k)})^{\gamma_k} * (t)^{\gamma_k}$, where $\mathbf{Q}$ is the vector of all covariates of interest in this order $(1, Z, X1, X2)$, which then correspond to the cause specific vector of coefficients $\zeta_k = (\zeta_0, \zeta_z, \zeta_{x1}, \zeta_{x2})$. The overall survivor function is then given by: $Surv(t|\mathbf{Q}) = \text{Exp}(-\sum_k H_k(t|\mathbf{Q}))$.

We create overall survival times by inverting the CDF, one less the survivor, using the probability integral transform to obtain overall survival times, $Tov$. We then determine which of the event types a time belongs to by randomly generating from a Bernoulli with probability $h_m(Tov|\mathbf{Q})/(h_m(Tov|\mathbf{Q}) + h_{m'}(Tov|\mathbf{Q}))$ and assigning event type 1 if 1 and 2 if 0. We then generate censoring times using `rweibull` with shape parameter equal to $e^{\mathbf{Q}^T\alpha}$ and scale parameter $2\gamma_1$. The intercept (i.e., first element) of $\alpha$ determines the amount of censoring, and whether the remaining coefficients are non-zero determines whether the censoring depends on covariates.

The true values of the coefficients were determined by generated a very large sample of covariates $\mathbf{Q}$, then calculating the corresponding true values of the cumulative incidence or restricted mean life time lost, and finally regressing those true values against the covariates using the link function. Samples large enough to acheive a precision of 1e-4 on the coefficient values were used.

# Data example

## Model estimation

## Additional features

Plotting of residuals Perme and Andersen [2008], Inverse probability weighted for case cohort studies Parner et al. [2020], Prediction,

# Conclusion

# References

Per Kragh Andersen and Maja Pohar Perme. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research*, 19(1):71–99, February 2010. ISSN 0962-2802, 1477-0334. doi: 10.1177/0962280209105020. URL http://journals.sagepub.com/doi/10.1177/0962280209105020.

Per Kragh Andersen, John P. Klein, and Susanne Rosthøj. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1):15–27, March 2003. ISSN 0006-3444. doi: 10.1093/biomet/90.1.15. URL https://academic.oup.com/biomet/article/90/1/15/218726.

Per Kragh Andersen, Vladimir Canudas-Romo, and Niels Keiding. Cause-specific measures of life years lost. *Demographic Research*, 29:1127–1152, December 2013. ISSN 1435-9871. doi: 10.4054/DemRes.2013.29.41. URL http://www.demographic-research.org/volumes/vol29/41/.

Jan Beyersmann, Aurelien Latouche, Anika Buchholz, and Martin Schumacher. Simulating competing risks data in survival analysis. *Statistics in medicine*, 28(6):956–971, 2009.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.

Thomas A. Gerds. *prodlim: Product-Limit Estimation for Censored Event History Analysis*, 2019. URL https://CRAN.R-project.org/package=prodlim. R package version 2019.11.13.

Louis A Jaeckel. *The infinitesimal jackknife*. Bell Telephone Laboratories, 1972.

Morten Overgaard, Erik Thorlund Parner, and Jan Pedersen. Pseudo-observations under covariate-dependent censoring. *Journal of Statistical Planning and Inference*, 202:112–122, 2019.

Erik T Parner, Per K Andersen, and Morten Overgaard. Cumulative risk regression in case–cohort studies using pseudo-observations. *Lifetime Data Analysis*, pages 1–20, 2020.

Maja Pohar Perme and Per Kragh Andersen. Checking hazard regression models using pseudo-observations. *Statistics in Medicine*, 27(25):5309–5328, 2008.

Terry M Therneau. *A Package for Survival Analysis in R*, 2020. URL https://CRAN.R-project.org/package=survival. R package version 3.1-12.