Stop the abuse: A plea for a more principled approach to the analysis of time-to-event endpoints with competing risks, with a focus on analysis of AEs

Kaspar Rufibach
Methods, Collaboration, and Outreach Group, Roche Basel
IBIG Journal Club, 23th October 2023

Acknowledgments

- SAVVY consortium, specifically Regina Stegherr, Jan Beyersmann, Claudia Schmoor, Tim Friede.
- Thomas Künzel.
- X-industry working group on estimands for time-to-event endpoints.
- Competing risks + estimands: Jan Beyersmann, Marcel Wolbers.
- Comments on linkedin post.

Take home messages

Need accurate estimates of P(AE) + comparison between arms.

IP and (1 - KM) biased irrespective of what we use them for.

Bias "does not cancel out" when comparing P(AE) between arms in RCT.

Let me explain.

Agenda

1. Take home messages
2. Estimation of P(AE)
   - The SAVVY project
   - Bias of common estimators of AE risk
   - Bias of common estimators of relative AE risk
3. Take home messages
4. Resources and future plans

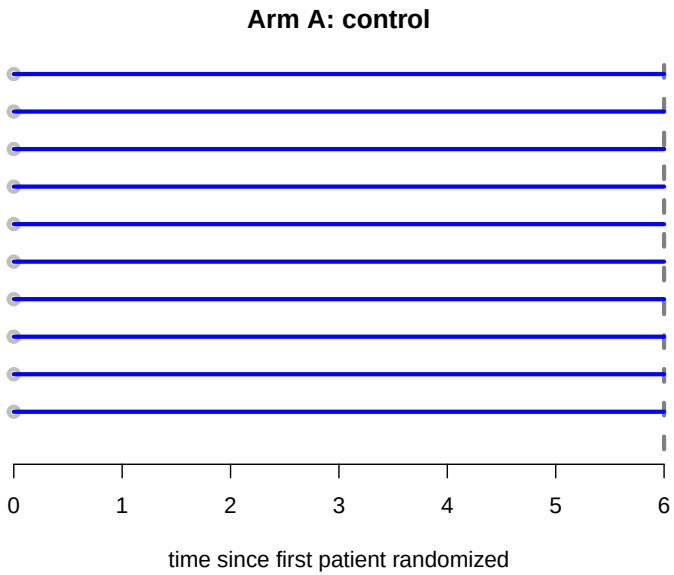
Assume you want to assess whether a new drug prolongs OS in an RCT with staggered recruitment.

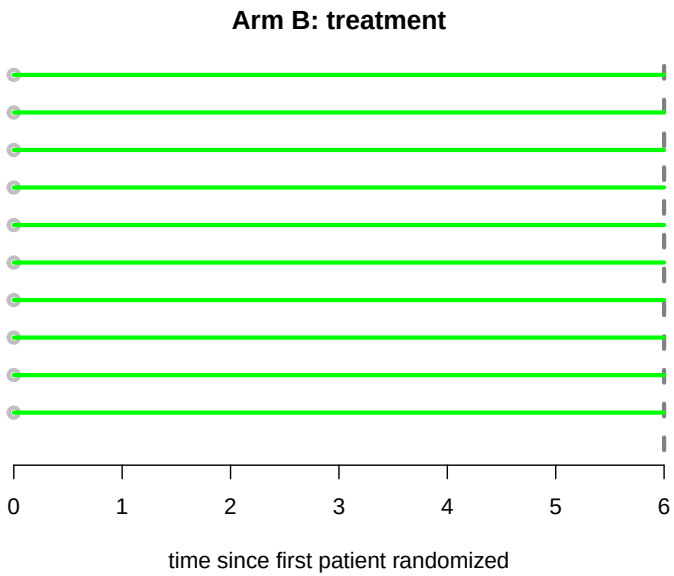
Clinicians proposal: cut data at four years and compare proportions of those who died.

What would you say?

Estimation of P(AE)

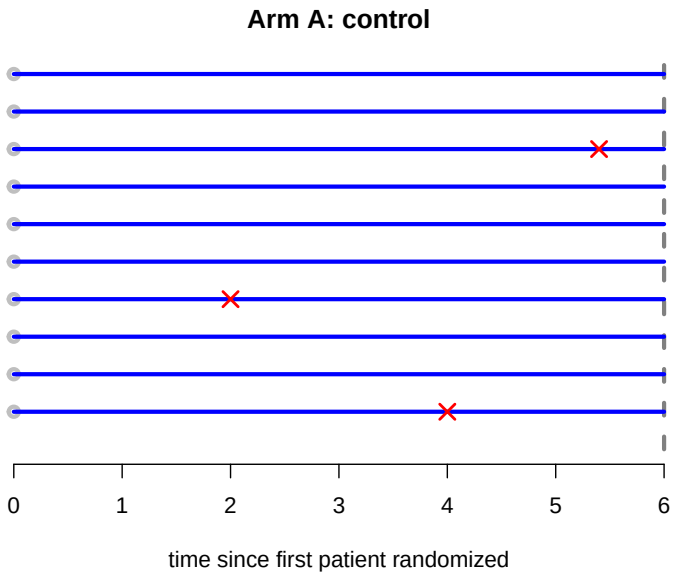
Estimation of P(AE)





- 2-arm RCT.
- 10 patients per arm.
- All patients randomized on same day.
- All patients observed for 6 months.

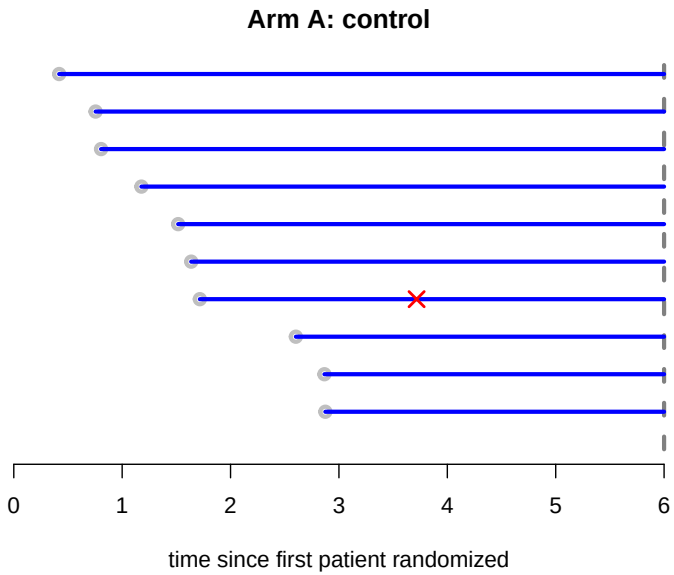
Estimation of P(AE)

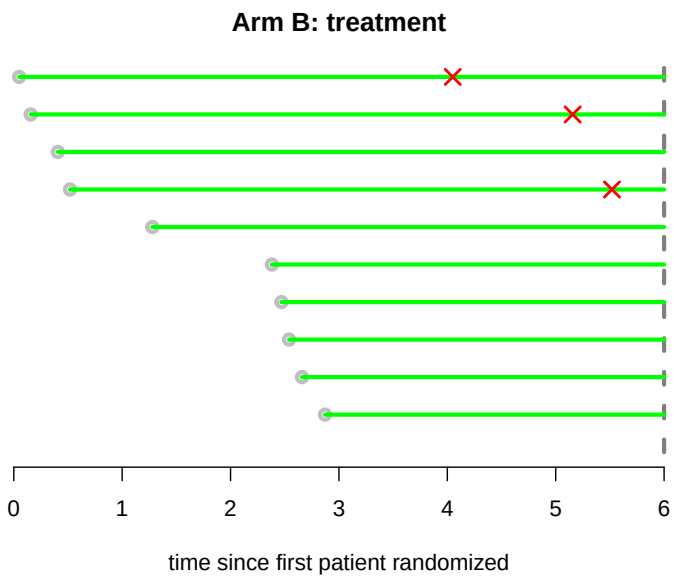


- 2-arm RCT.
- 10 patients per arm.
- All patients randomized on same day.
- All patients observed for 6 months.

$$\text{P(AE in A)} = 3 \,/\, 10 = 0.30,$$
$$\text{P(AE in B)} = 4 \,/\, 10 = 0.40.$$

Estimation of P(AE): staggered entry





- 2-arm RCT.
- 10 patients per arm.
- Patients enter the trial over time.
- All patients observed until cutoff.

P(AE in A) = 1 / 10 = 0.10,
P(AE in B) = 3 / 10 = 0.30.

Is this what we want?

Staggered entry / censoring only
removes AE events $\Rightarrow$ **underestimation**.

What do these proportions estimate?

Incidence proportion in experimental arm in interval from 0 to t:

$$\widehat{IP}_E(t) \quad = \quad \frac{\text{Number of patients with AE in } [0,t] \text{ and that this AE is observed}}{n_E}.$$

$\widehat{IP}_E(t)$ estimates:

P(AE happens in $[0,t]$ and that this AE is observed **before censoring**).

$\widehat{IP}_E(t) \leq \widehat{P}(\text{AE happens in } [0,t]) \Rightarrow \widehat{IP}_E(t)$ **underestimates** absolute AE risk.

With censoring it is unclear which quantity $\widehat{IP}_E$ is estimating.

Simple incidence proportion is biased
if we have unequal follow-up or censoring.

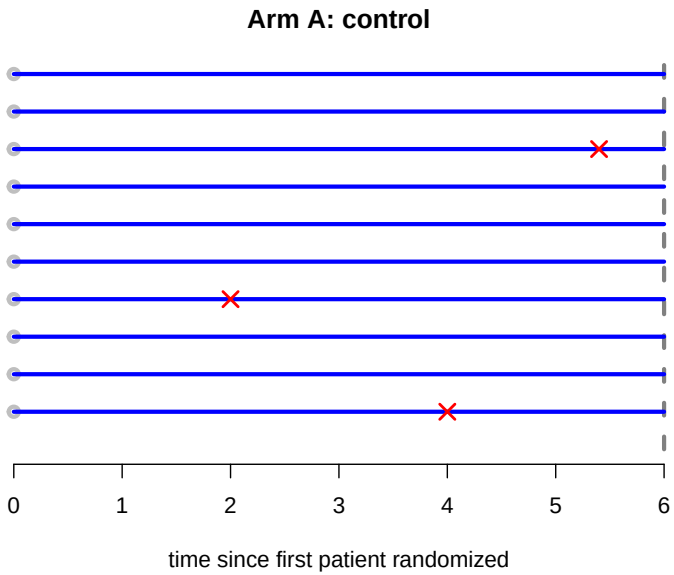
Estimate P(AE) using time-to-AE

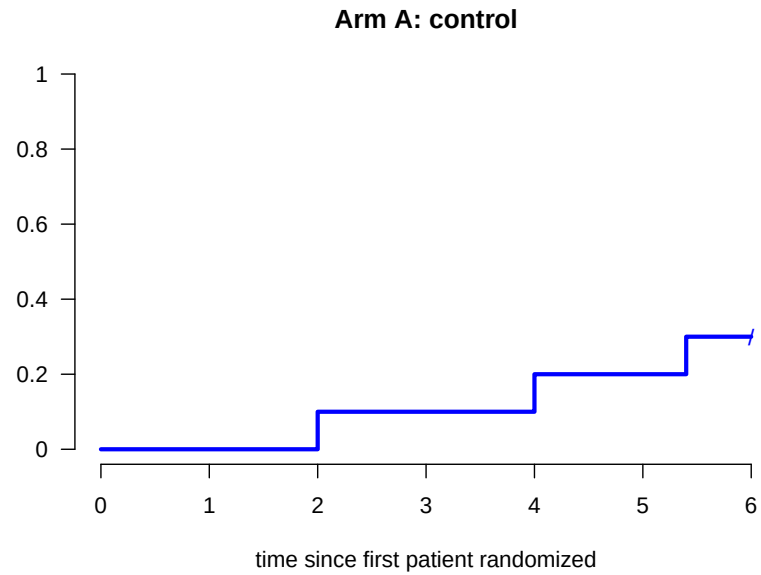
Consider time-to-first-AE

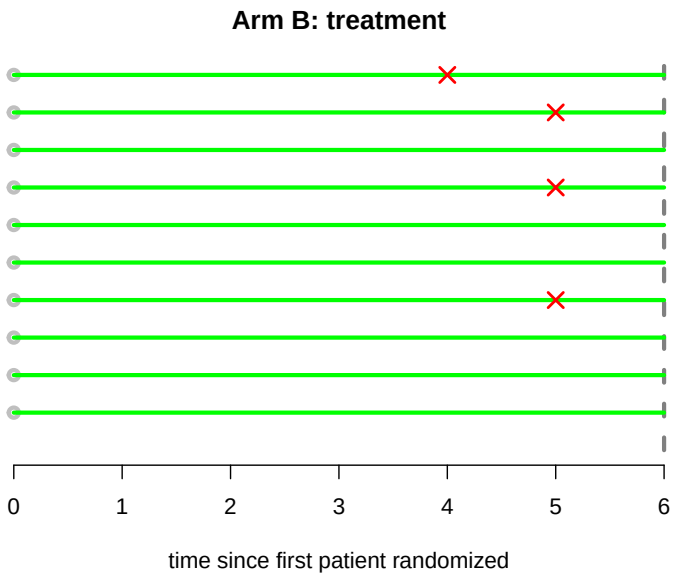
Redefine question: Consider **time-to-first-AE**.

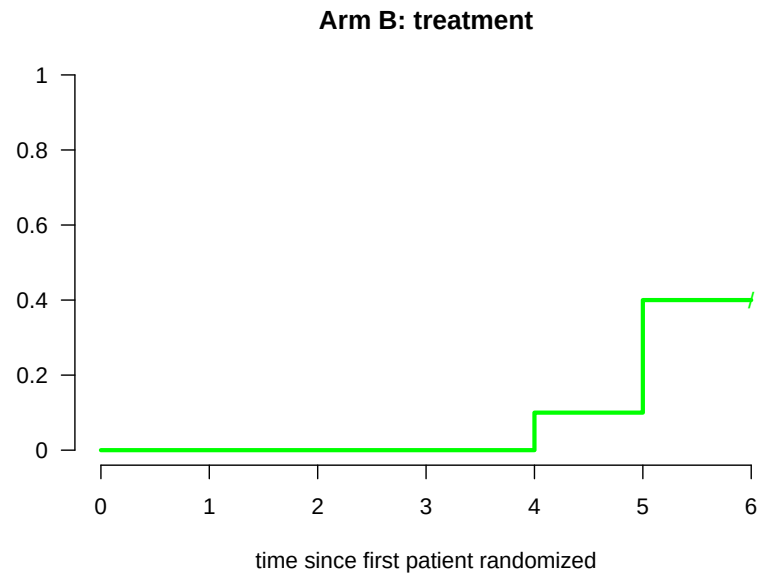
- Estimate P(AE happens in $[0, t]$) using 1 - Kaplan-Meier.
- Correctly accounts for **censoring**.
- Consistently estimates AE risk at t, accounting for varying follow-up.

Estimation of P(AE)

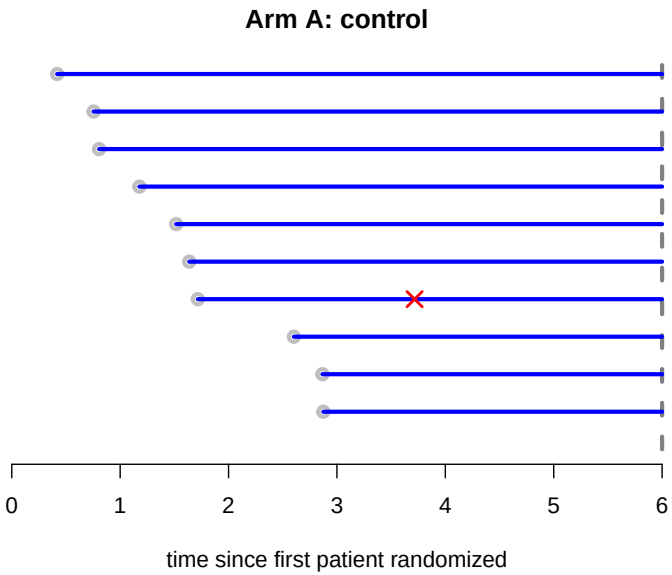


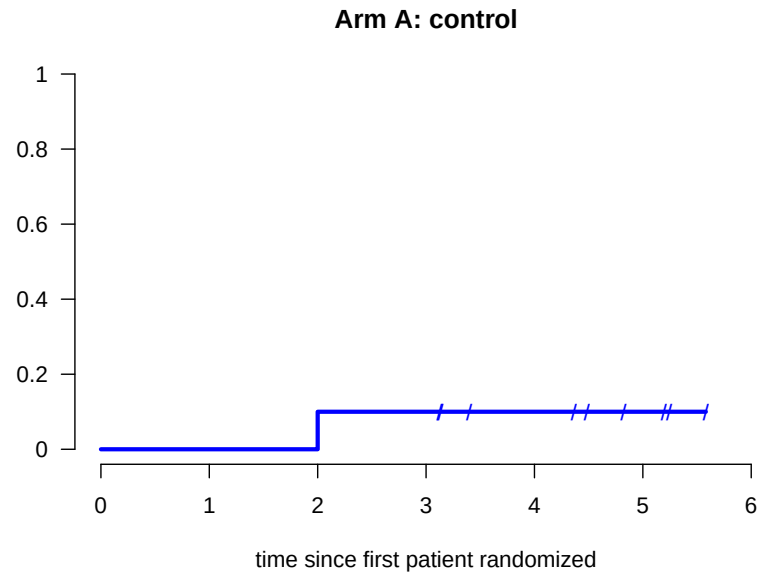


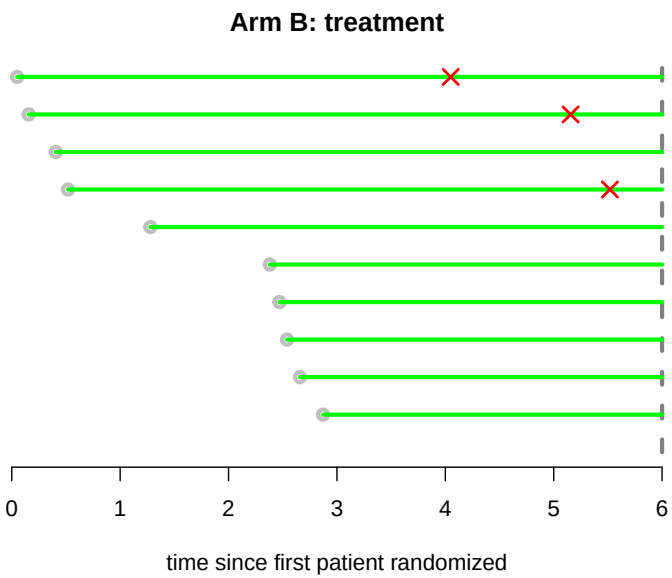


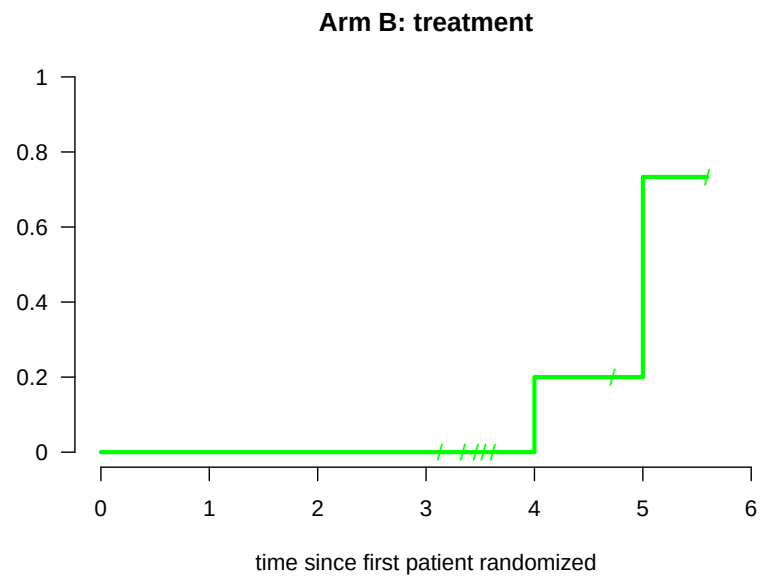


Estimation of P(AE): staggered entry





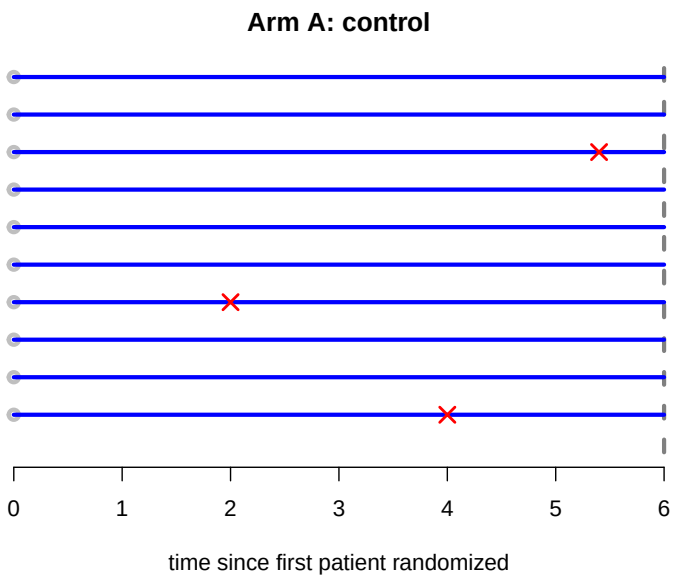


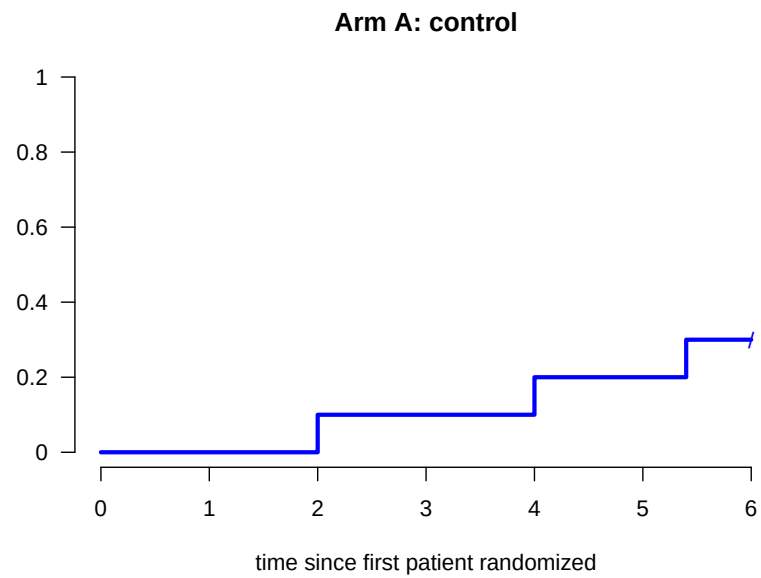


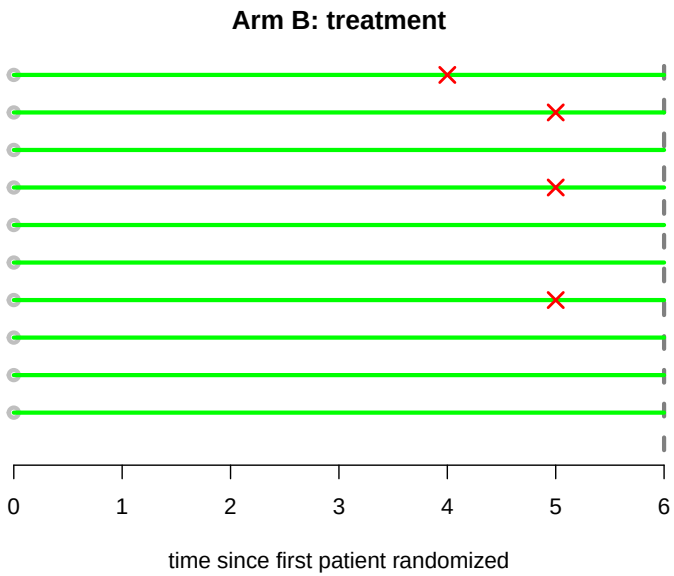
Competing events

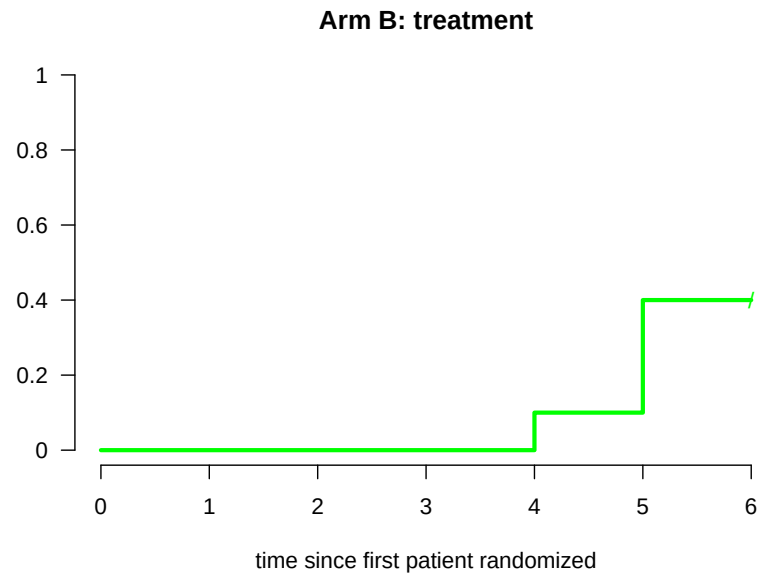
(= competing risk)

Estimation of P(AE)

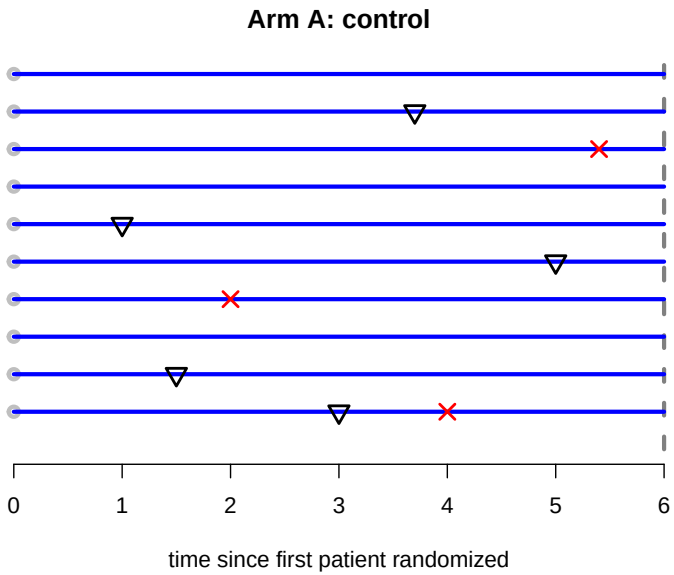


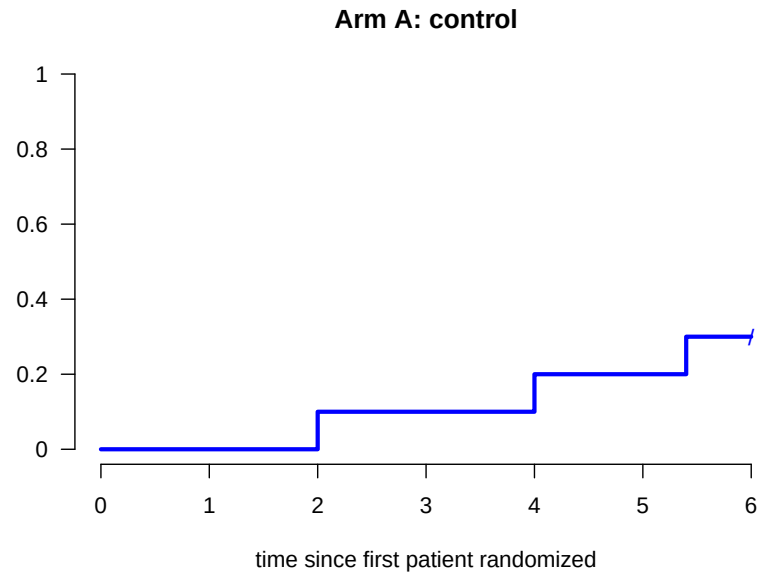


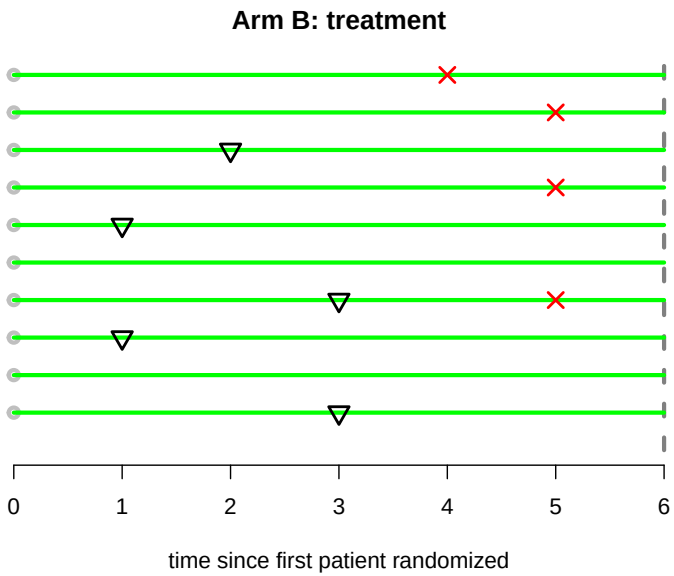


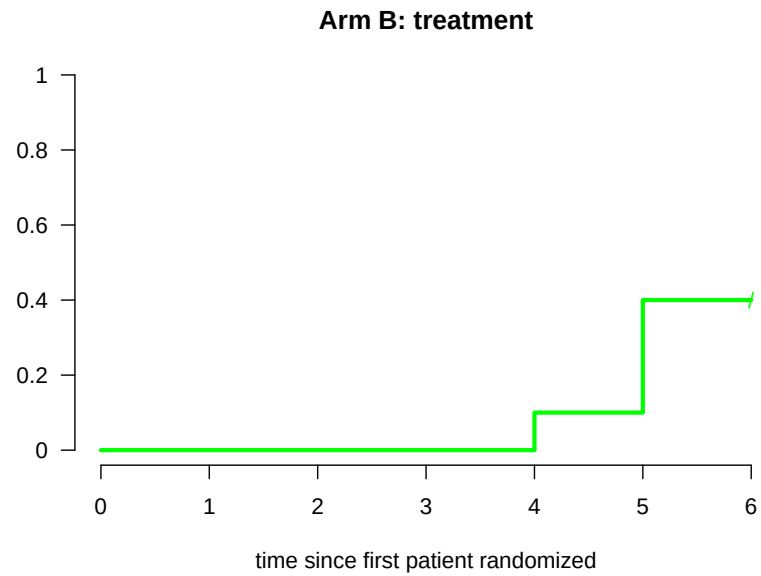


Estimation of P(AE): competing event of death

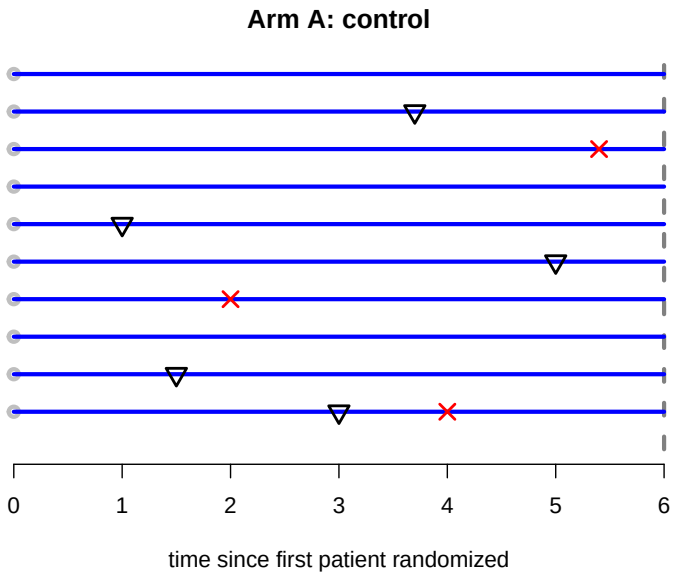


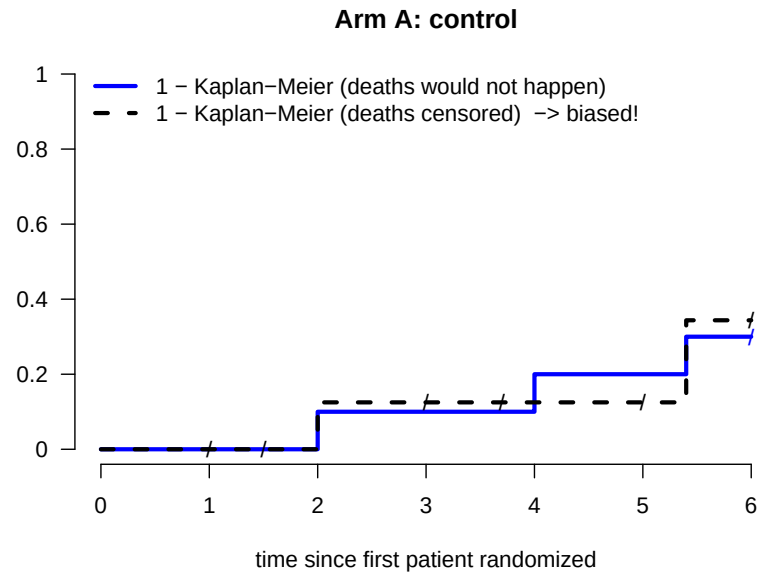


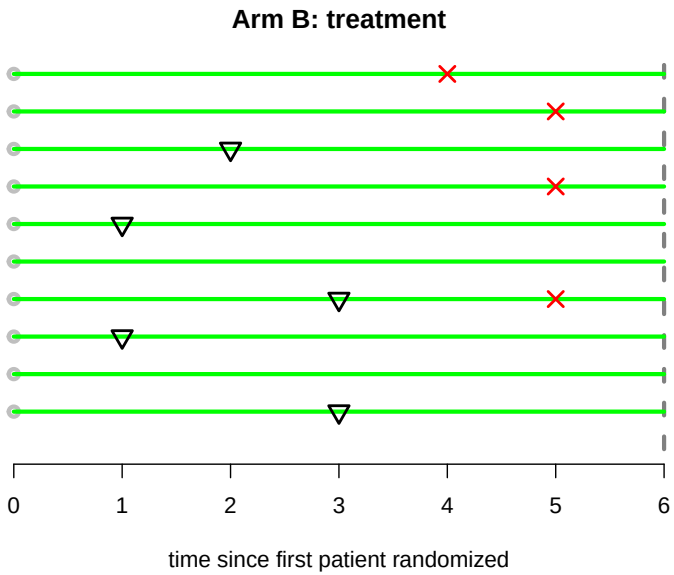


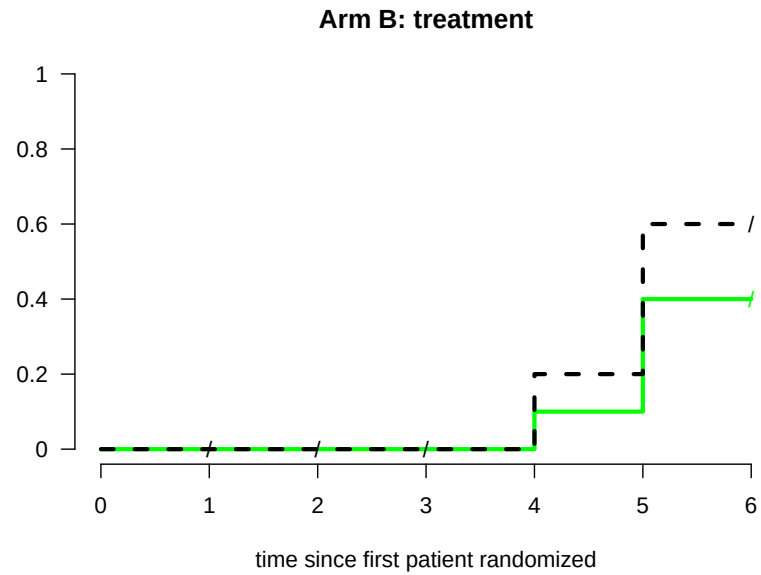


Estimation of P(AE): competing event of death









What does (1 - $\widehat{\text{KM}}$) with censoring of CEs estimate?

Administrative censoring: patients may still experience event at later time point.

Not for CEs!

What does (1 - $\widehat{\text{KM}}$) with censoring of CEs estimate?

- **Violates independent censoring assumption**:
  - Patient censored at death will NEVER experience AE.
  - Patients who will never experience AE treated as if they could still have one.
- Less than 100% of patients experience AE **before** death:
  - Some die before AE $\Rightarrow$ P(AE) < 1.
  - But (1 - $\widehat{\text{KM}}$) approaches 1 $\Rightarrow$ naive (1 - $\widehat{\text{KM}}$) **overestimates** P(AE).

Abandon!

Although tutorial articles are available, too many studies are susceptible to competing risk bias which **can be avoided by using adequate statistical methodology**. *There is* **no excuse not to use it**, *and Kaplan-Meier methodology should be completely abandoned in the analysis of end points with competing risks in all journals.*

Schumacher et al. (2016)

1 - Kaplan-Meier is biased if we have competing events.

Is this relevant at all?

How large can the bias be?

The SAVVY project

The SAVVY project

Survival analysis for **A**d**V**erse events with **V**ar**Y**ing follow-up times:

Goal: improve analyses of AE data in clinical trials through use of **survival techniques** appropriately dealing with

- varying follow-up times,
- censoring,
- competing events.

SAVVY webpage

9 pharma


MERCK
Boehringer Ingelheim
Bristol Myers Squibb
Roche
Lilly
janssen
Pfizer
NOVARTIS
BAYER

9 pharma + 3 universities

The SAVVY project

Federated learning: central analysis team:

- Developed macros (R + SAS). Validated R package under development.
- Every sponsor ran them on their data.
- Only share aggregated data.
- Central team performed meta-analysis.

Data from **17 RCTs** in various indications.

200 - 7171 patients.

186 AEs: selected by sponsor.

The SAVVY project

Estimate P(AE) at latest available follow-up with various estimators:

- Estimate **P(AE) in one arm** (the experimental).
- Estimate **relative risk** in RCTs using risk and hazard ratio.

CEs in SAVVY:

- **Hard**: Death - AE after death impossible.
- **Soft**: lost to follow-up, withdrawal of consent, treatment discontinuation $\Rightarrow$ AE of interest can in principle still occur but is not observed due to end of follow-up.

Interest in estimation of P(AE), not in P(specific CE) $\Rightarrow$ lump all CEs together, not interested in cumulative incidence of CE.

Goal: compare bias of estimators.

What is "gold standard"?

Gold standard: Aalen-Johansen estimator

SAVVY: **Empirical** bias evaluation within RCTs.

What is "best" estimator to benchmark against?

Estimator	Accounts for censoring	Accounts for CEs
Incidence proportion	No	Yes
1 - Kaplan-Meier	Yes	No
Aalen-Johansen estimator	Yes	Yes

All **nonparametric**: no constant hazard assumption.

Aalen-Johansen:

- Generalizes Kaplan-Meier to competing risk and general multistate models.
- **No censoring**: Aalen-Johansen = incidence proportion.
- **No competing events**: Aalen-Johansen = (1 - Kaplan-Meier).

Bias of common estimators of AE risk

Estimation of AE risk: incidence proportion

Experimental arm.

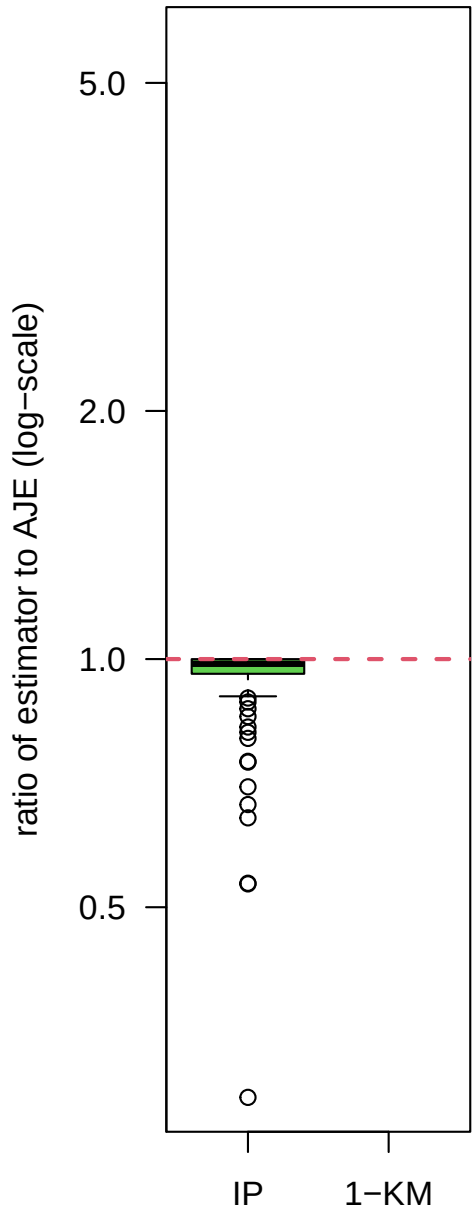
Evaluated at maximal observed follow-up time τ.

Incidence proportion:

- Accounts for CEs but not censoring.
- Point in boxplot: corresponds to ratio of $\widehat{IP}_E(\tau)$ to gold standard for given AE.
- Ratio $= 1$: $\widehat{IP}_E(\tau)$ gives same AE risk estimate as gold standard.
- **Underestimation of P(AE) up to factor THREE!**

Overall performance not too bad. Why?

Datasets have many **soft** CEs $\Rightarrow$ little censoring.



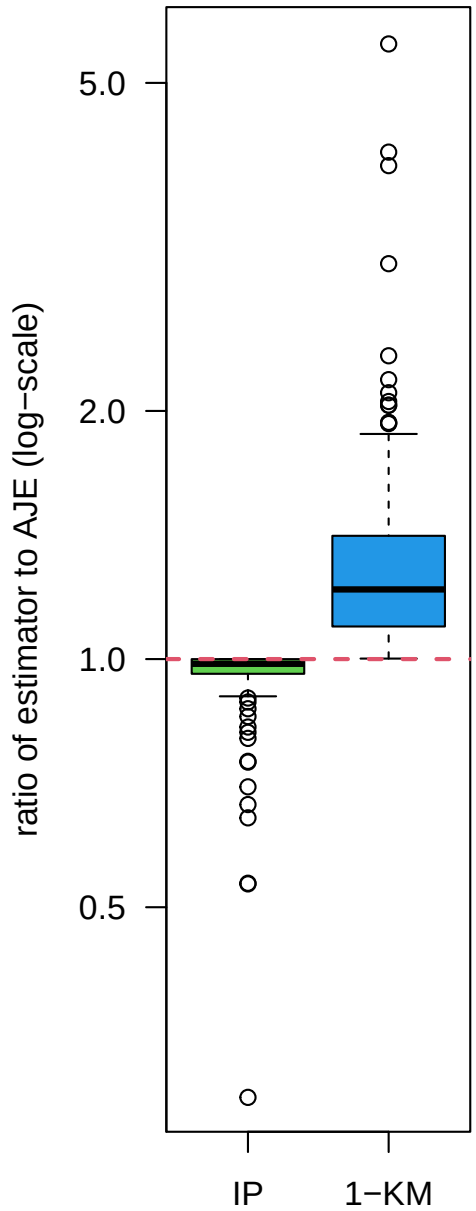
Estimation of AE risk: 1 - Kaplan-Meier

Experimental arm.

Evaluated at maximal observed follow-up time τ.

1 - Kaplan-Meier:

- Accounts for censoring but not CEs.
- Point in boxplot: corresponds to ratio of (1 - $\widehat{KM})_E(\tau)$ to gold standard for given AE.
- Ratio = 1: (1 - $\widehat{KM})_E(\tau)$ gives same AE risk estimate as gold standard.
- **Overestimation of P(AE) up to factor FIVE!**



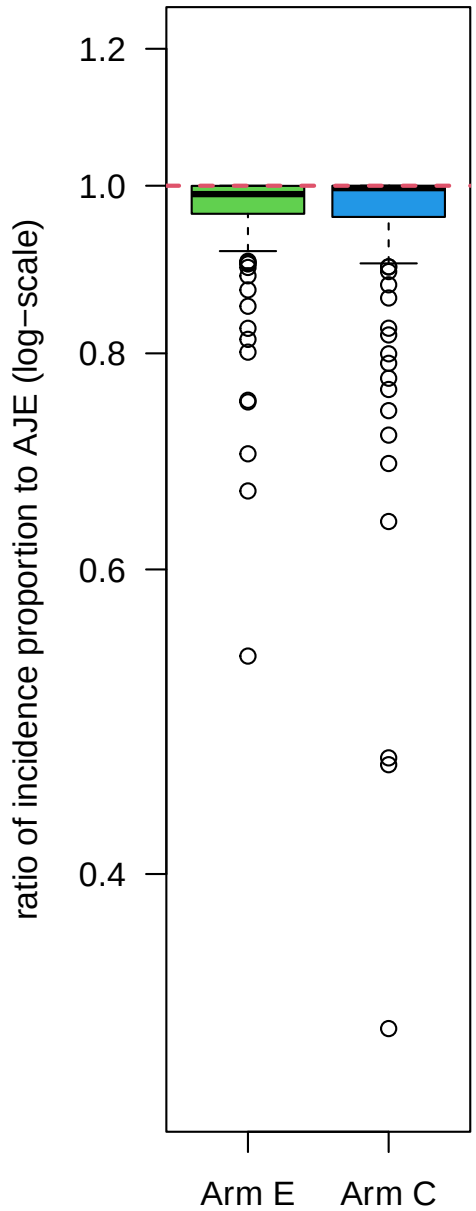
Bias of common estimators of relative AE risk

Estimation of relative AE risk: incidence proportion

Evaluated at minimum of maximal observed follow-up τ.

Incidence proportion:

- Point in boxplot: corresponds to ratio of $\widehat{IP}(\tau)$ to gold standard for given AE and treatment arm.
- Ratio $= 1$: $\widehat{IP}(\tau)$ gives same AE risk estimate as gold standard.
- **Underestimation** of P(AE) compared to gold standard.

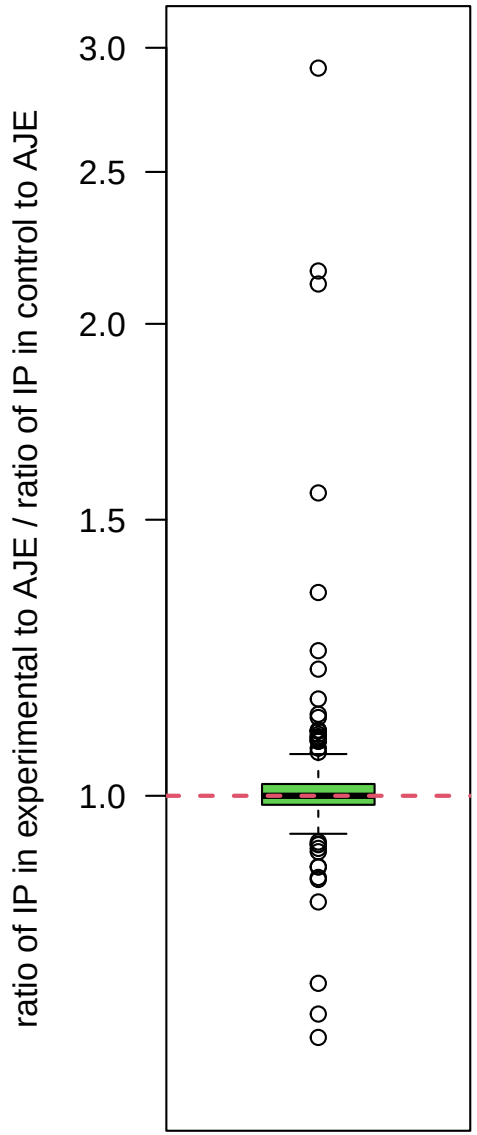


Estimation of relative AE risk: incidence proportion

Evaluated at minimum of maximal observed follow-up τ.

Incidence proportion:

- Point in boxplot: corresponds to ratio of $\widehat{IP}_E(\tau)/\widehat{IP}_C(\tau)$ to gold standard for given relative AE risk.
- Ratio $= 1$: $\widehat{IP}_E(\tau)/\widehat{IP}_C(\tau)$ gives same relative AE risk estimate as gold standard.
- Over- and underestimation observed.
- **Overestimation of RR up to factor of almost 3.**

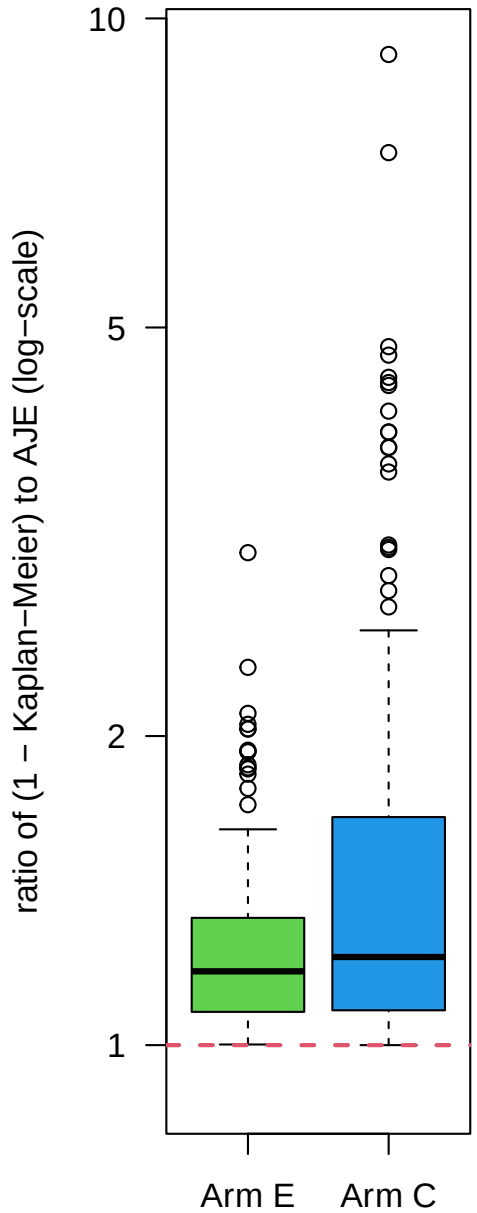


Estimation of relative AE risk: (1 - KM)

Evaluated at minimum of maximal observed follow-up τ.

1 - Kaplan-Meier:

- Point in boxplot: corresponds to ratio of (1 - $\widehat{KM}$)(τ) to gold standard for given AE and treatment arm.
- Ratio = 1: (1 - $\widehat{KM}$)(τ) gives same AE risk estimate as gold standard.
- **Overestimation** of P(AE) compared to gold standard.

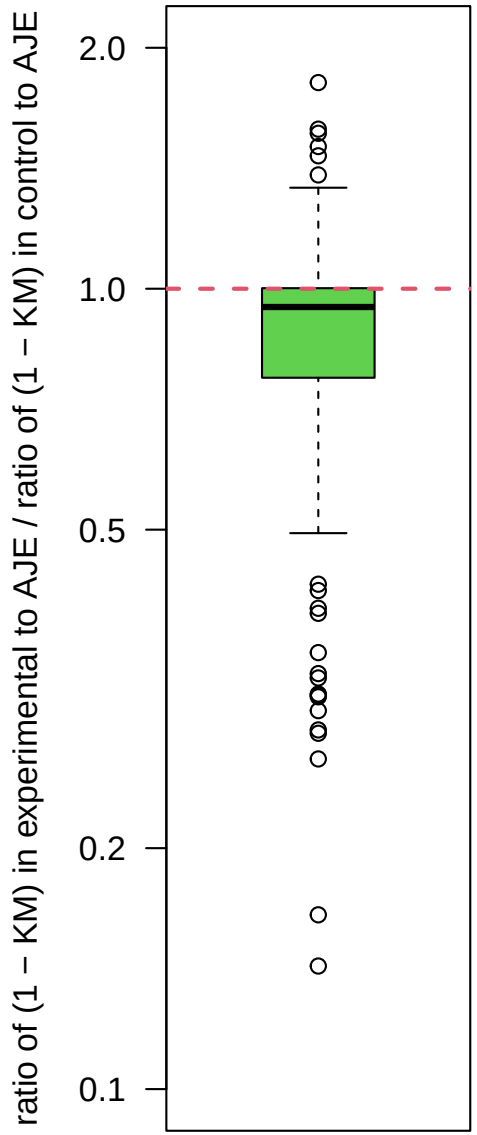


Estimation of relative AE risk: (1 - KM)

Evaluated at minimum of maximal observed follow-up τ.

1 - Kaplan-Meier:

- Point in boxplot: corresponds to ratio of (1 - $\widehat{KM})_E(\tau)$ / (1 - $\widehat{KM})_C(\tau)$ to gold standard for given AE.
- Ratio = 1: (1 - $\widehat{KM})_E(\tau)$ / (1 - $\widehat{KM})_C(\tau)$ gives same relative AE risk estimate as gold standard.
- Over- and underestimation observed.
- **Underestimation of RR up to factor of >4.**



Arm-wise bias does not cancel out in relative comparisons.

Now we have seen what does not work.

But what does work?

Aalen-Johansen: properly accounts for varying follow-up times and competing risks.

Before you ask...

Before you ask...

Focus on bias - what about variability?

- Focus today with IP rarely on variability either!
- Simulation study for 2-arm comparisons: Stegherr et al.(2021c).

We do not collect data necessary to estimate P(AE) with AJE?

- ICH E9(R1) estimands addendum: **clinical trial objective** dictates data collection and analytical method!
- Clarify **clinical trial objective** also for analysis of safety!
- **Proper definition of CE** requires understanding and discussion of therapeutic area.

Before you ask...

Does normalization by exposure time not solve the problem?

- **Incidence density**. See backup for details.
- A priori estimates **AE hazard**, not P(AE). Can be turned into estimator of P(AE).
- Assumes **exponentiality** of AE hazard.
- Incidence density for each CE.

Can we use IP for "signal detection" or other purposes?

Biases = statistical properties of IP, (1 - KM).

Independent of what we use estimates of P(AE) for!

Take home messages

Need accurate estimates of P(AE) + comparison between arms.

IP and (1 - KM) biased irrespective of what we use them for.

Bias "does not cancel out" when comparing P(AE) between arms in RCT.

Resources and future plans

Resources

SAVVY webpage:

- Exemplary code for all methods.
- All papers and talks.
- Papers:
  - SAP: Stegherr et al. (2021a).
  - Methods: Stegherr et al.(2021c).
  - 1-sample: Stegherr et al. (2021b).
  - 2-sample: Rufibach et al. (2022).
- Effective statistician podcasts:
  - About SAVVY: https://theeffectivestatistician.com/the-analysis-of-adverse-events-done-right-savvy/.
  - 200th episode with 10% most downloaded podcasts: https://theeffectivestatistician.com/200th-episode/.

Slides available on www.kasparrufibach.ch.

Future plans

Estimate **disease-specific P(AE)'s**, properly discussing therapeutic area specific CEs.

Influence **updating of guidelines**.

Thank you for your attention.

kaspar.rufibach@roche.com

Slides can be downloaded on

www.kasparrufibach.ch

References I

- Rufibach, K., Stegherr, R., Schmoor, C., Jehl, V., Allignol, A., Boeckenhoff, A., Dunger-Baldauf, C., Eisele, L., Künzel, T., Kupas, K., Friedhelm, L., Trampisch, M., Zhao, Y., Friede, T. and Beyersmann, J. (2022). Survival analysis for AdVerse events with VarYing follow-up times (SAVVY) – comparison of adverse event risks in randomized controlled trials. *Statistics in Biopharmaceutical Research, accepted* . `https://arxiv.org/abs/2008.07881`
- Schumacher, M., Ohneberg, K. and Beyersmann, J. (2016). Competing risk bias was common in a prominent medical journal. *Journal of clinical epidemiology* **80** 135–136.
- Stegherr, R., Beyersmann, J., Jehl, V., Rufibach, K., Leverkus, F., Schmoor, C. and Friede, T. (2021a). Survival analysis for adverse events with varying follow-up times (savvy): Rationale and statistical concept of a meta-analytic study. *Biometrical journal. Biometrische Zeitschrift* **63** 650–670.
- Stegherr, R., Schmoor, C., Beyersmann, J., Rufibach, K., Jehl, V., Brückner, A., Eisele, L., Künzel, T., Kupas, K., Langer, F., Leverkus, F., Loos, A., Norenberg, C., Voss, F. and Friede, T. (2021b). Survival analysis for AdVerse events with VarYing follow-up times (SAVVY)-estimation of adverse event risks. *Trials* **22** 420.
- Stegherr, R., Schmoor, C., Lübbert, M., Friede, T. and Beyersmann, J. (2021c). Estimating and comparing adverse event probabilities in the presence of varying follow-up times and competing events. *Pharm Stat* **20** 1125–1146.

Doing now what patients need next

R version and packages used to generate these slides:

R version: R version 4.2.3 (2023-03-15 ucrt)

Base packages: stats / graphics / grDevices / utils / datasets / methods / base

Other packages: ggplot2 / etm / cmprsk / mvna / prodlim / survival / reporttools / xtable

This document was generated on 2023-10-12 at 22:15:10.