Stop the abuse: A plea for a more principled approach to the analysis of adverse events

Kaspar Rufibach
Methods, Collaboration, and Outreach Group, Roche Basel
AEs in CTs Sub-group of TMRP Outcomes Working Group, 6 May 2022

Agenda

1. Estimation of P(AE)
2. The SAVVY project
3. Conclusions
4. Resources
5. The SAVVY project
6. Bias of common estimators of AE risk
7. Bias of common estimators of relative AE risk
8. Statistical inference for competing risks

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?

If multiple AEs / patient
$\Rightarrow$ only interested in 1st

Varying follow-up time

Typical AE reporting

Category	All Adverse Events	
	Obinutuzumab Group (N=595)	Rituximab Group (N=597)
Infection*	460 (77.3)	418 (70.0)
Neutropenia	301 (50.6)	269 (45.1)
Infusion-related event†		
Any event	406 (68.2)	349 (58.5)
Antibody-related event	353 (59.3)	292 (48.9)
Tumor lysis syndrome	6 (1.0)	3 (0.5)
Cardiac event‡	78 (13.1)	58 (9.7)
Thrombocytopenia	68 (11.4)	45 (7.5)
Second neoplasm§	43 (7.2)	30 (5.0)
Nonmelanoma skin cancer	18 (3.0)	14 (2.3)
Hematologic event¶	6 (1.0)	0
Other	22 (3.7)	18 (3.0)
Myelodysplastic syndrome	2 (0.3)	0
Gastrointestinal perforation	4 (0.7)	3 (0.5)
Hemorrhagic event	57 (9.6)	62 (10.4)

Marcus *et al.* (2017)

What do we use these tables for?

What conclusions do we draw?

AE where accurate estimate of P(AE) is relevant

Category	All Adverse Events	
	Obinutuzumab Group (N=595)	Rituximab Group (N=597)
Infection*	460 (77.3)	418 (70.0)
Neutropenia	301 (50.6)	269 (45.1)
Infusion-related event†		
Any event	406 (68.2)	349 (58.5)
Antibody-related event	353 (59.3)	292 (48.9)
Tumor lysis syndrome	6 (1.0)	3 (0.5)
Cardiac event‡	78 (13.1)	58 (9.7)
Thrombocytopenia	68 (11.4)	45 (7.5)
Second neoplasm§	43 (7.2)	30 (5.0)
Nonmelanoma skin cancer	18 (3.0)	14 (2.3)
Hematologic event¶	6 (1.0)	0
Other	22 (3.7)	18 (3.0)
Myelodysplastic syndrome	2 (0.3)	0
Gastrointestinal perforation	4 (0.7)	3 (0.5)
Hemorrhagic event	57 (9.6)	62 (10.4)

Marcus *et al.* (2017)

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**).

$\widehat{IP}_E(t) \leq \widehat{P}$(experiencing the AE) $\Rightarrow \widehat{IP}_E(t)$ **underestimates** absolute AE risk.

What does the incidence proportion estimate?

Incidence proportion: Estimates P(AE) within a given amount of follow-up.

If we have

- **censoring**: unclear what $\widehat{IP}_E(t)$ is estimating.
- **treatment effect**: comparisons between groups difficult, if not unfair.

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

Implementing estimation of P(AE)

Methodology is available: Aalen-Johansen estimator.

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

Proper definition of CE requires understanding of TA and discussion.

Is this relevant at all?

How large can the bias be?

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.

Nine pharmaceutical companies: F. Hoffmann-La Roche, Novartis, Janssen-Cilag, Bristol-Myers-Squibb, Lilly, Pfizer, Merck, Bayer, Boehringer Ingelheim.

One academic trial center: Freiburg.

University of Ulm & Göttingen.

The SAVVY project

Data from **17 RCTs** in various indications. 200 - 7171 patients. 186 AEs.

Avoid issues with data sharing $\Rightarrow$ central analysis team:

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

Before you ask...

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

- ICH E9(R1) estimands addendum: **scientific objective** dictates data collection and analytical method!
- Clarify **scientific objective** also for analysis of safety!
  - First AE? Recurrent AE?
  - Safety: we "simply use" IP, but what is the scientific objective?
  - Discuss CEs.

Does normalization by exposure time not solve the problem, e.g. incidence density? No!

Biases are **appreciable**!

Guidelines need updates!

ICH:

- Methods to analyze safety data: E1, E2, E3, or E9.
- Describe analysis methods, primarilyincidence proportion, incidence density.
- Lack clear formulation of scientic objective.
- E9: Explicitly asks for "...appropriate use of **survival analysis** methods to exploit the potential relationship of the incidence of adverse events to duration of exposure and/or follow-up."

EMA anticancer guideline:

- Kaplan-Meier analysis of selected AEs, which considers censoring of events, may be useful. What about **competing risks**?
- IQWiG comment to earlier version: "Even if censoring due to a CE does not lead to different follow-up times the presence of CEs is still a problem, because the usual Kaplan-Meier method leads to biased estimations of absolute risks."

CONSORT harm: "Absolute risks for binary events per arm and per type and grade (follow-up/exposure time is differential and not comparable for all participants)." to be estimated using Kaplan-Meier.

Conclusions

- Safety likely to become more important for **differentiation** in competitive spaces. Need **accurate estimate of P(AE)**!
- What if AE of interest were **primary endpoint**? Would you use same analysis?
- IP and (1 - KM) **biased** irrespective of what we use them for ("signal detection" vs. "estimation of P(AE)".)
- Comparing AE risk between two arms in RCT: differences between estimators become more emphasized.
- Good understanding of statistical theory and properties. **No reason not to use CEs and MSMs**!
- SAVVY: template for cross-company collaboration without need for data sharing.

Outloock:

- Estimate disease-specific P(AE).
- Work towards updating guidelines.

Resources

SAVVY:

- Markdown with exemplary code for all methods: `https://numbersman77.github.io/AEprobs/SAVVY_AEprobs.html`.
- SAP: Stegherr *et al.* (2021c).
- Methods: Stegherr *et al.* (2021a).
- 1-sample: Stegherr *et al.* (2021b).
- 2-sample: Rufibach *et al.* (2020).

Effective statistician podcast: `https://theeffectivestatistician.com/the-analysis-of-adverse-events-done-right-savvy/`

Thank you for your attention.

kaspar.rufibach@roche.com

http://www.kasparrufibach.ch

numbersman77

numbersman77

References I

Marcus, R., Davies, A., Ando, K., Klapper, W., Opat, S., Owen, C., Phillips, E., Sangha, R., Schlag, R., Seymour, J. F., Townsend, W., Trneny, M., Wenger, M., Fingerle-Rowson, G., Rufibach, K., Moore, T., Herold, M., and Hiddemann, W. (2017). Obinutuzumab for the First-Line Treatment of Follicular Lymphoma. *N. Engl. J. Med.*, **377**(14), 1331–1344.

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. (2020). Survival analysis for AdVerse events with VarYing follow-up times (SAVVY) – comparison of adverse event risks in randomized controlled trials. *Submitted*.

Stegherr, R., Schmoor, C., LÃŒbbert, M., Friede, T., and Beyersmann, J. (2021a). Estimating and comparing adverse event probabilities in the presence of varying follow-up times and competing events. *Pharm Stat*, **20**(6), 1125–1146.

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**(1), 420.

References II

Stegherr, R., Beyersmann, J., Jehl, V., Rufibach, K., Leverkus, F., Schmoor, C., and Friede, T. (2021c). 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.

Backup

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.

Competing events

Events that preclude occurrence of AE of interest in time-to-first-event setting

What does (1 - KM) in presence of competing events estimate?

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

Estimand corresponding to (1 - KM) in presence of CEs:

- **Violates independent censoring assumption**: Patient censored from CE will NOT experience event of interest $\Rightarrow$ patients that will never fail treated as if they could still fail (they are censored).
- Less than 100% of patients experience AE **before** CE - some experience CE earlier! But (1 - KM) approaches 1 $\Rightarrow$ naive (1 - KM) **overestimates** P(AE).

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

Is this relevant at all?

How large can the bias be?

The SAVVY project

Survival analysis for **Ad****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.

Nine pharmaceutical companies: F. Hoffmann-La Roche, Novartis, Janssen-Cilag, Bristol-Myers-Squibb, Lilly, Pfizer, Merck, Bayer, Boehringer Ingelheim.

One academic trial center: Freiburg.

University of Ulm & Göttingen.

The SAVVY project

Data from **17 RCTs** in various indications. 200 - 7171 patients. 186 AEs.

Avoid issues with data sharing $\Rightarrow$ central analysis team:

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

Goal: compare bias of estimators.

What is "gold standard"?

Gold standard

Truth: simulation study for 2-arm comparisons: Stegherr *et al.* (2021a).

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

What is the "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.

Estimation of AE risk: incidence proportion

Experimental arm.

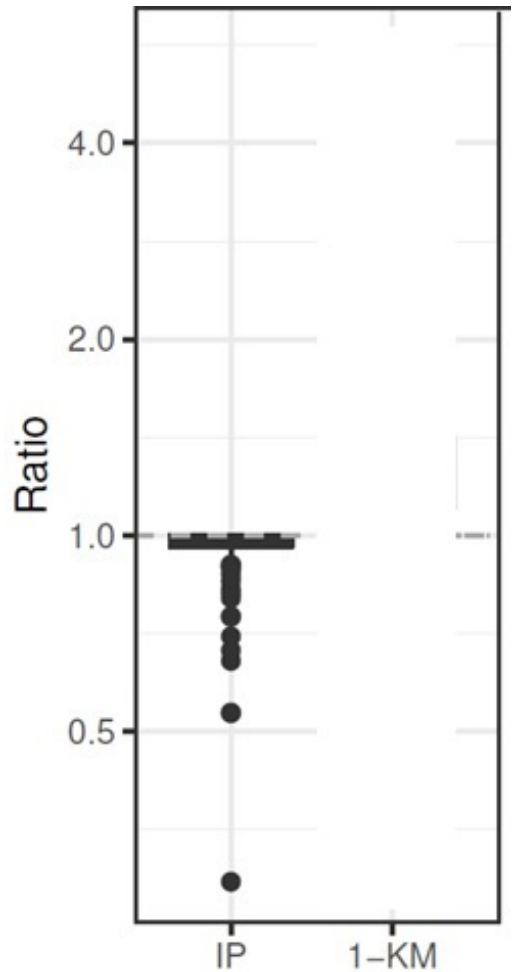
Evaluated at maximal observed follow-up time τ.

Incidence proportion:

- $\widehat{IP}_E(\tau)$ 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 CEs $\Rightarrow$ little censoring.



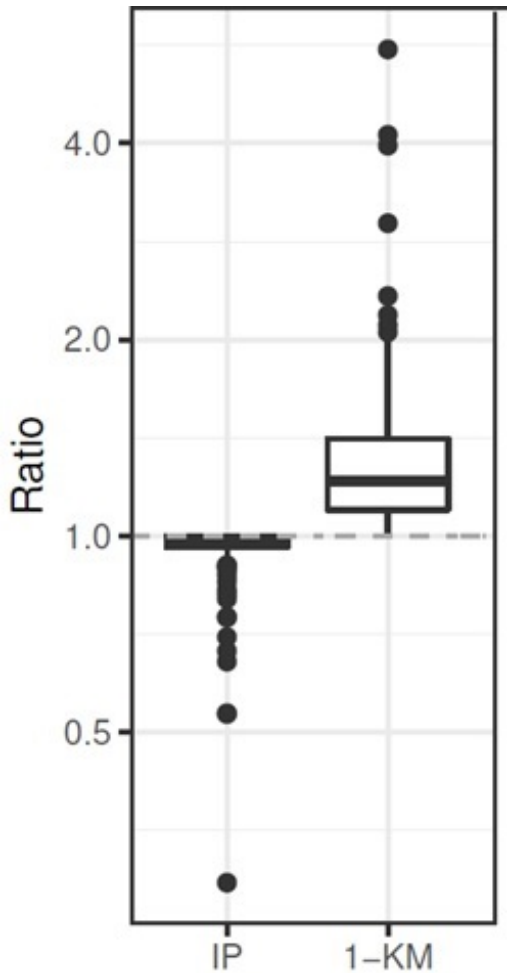
Estimation of AE risk: 1 - Kaplan-Meier

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!**

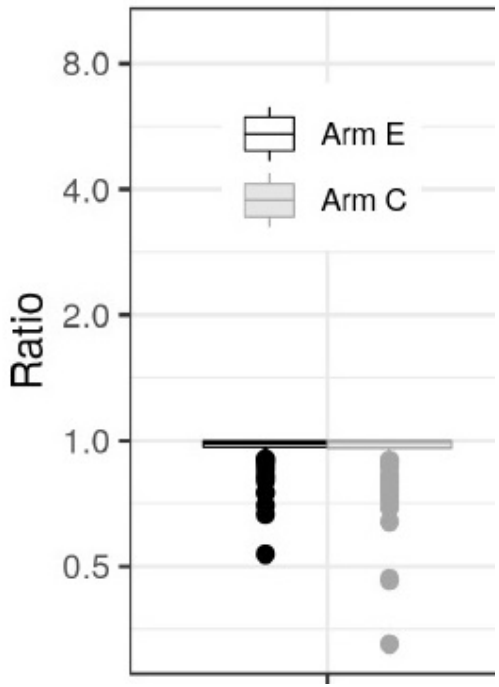


Estimation of relative AE risk: incidence proportion

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.

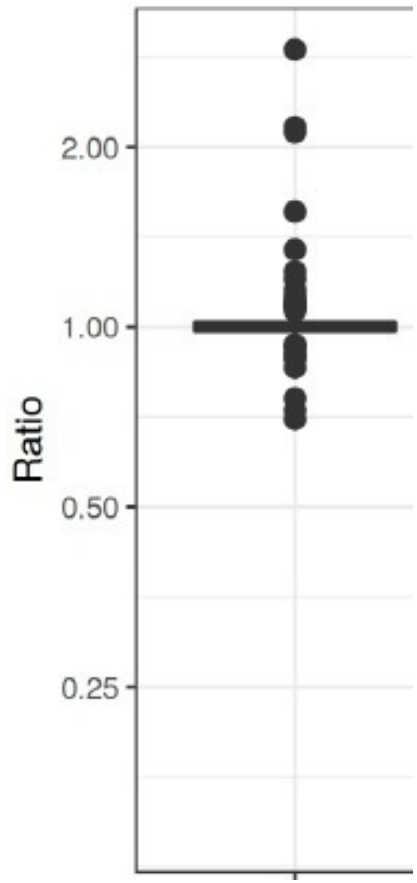


Estimation of relative AE risk: incidence proportion

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.**

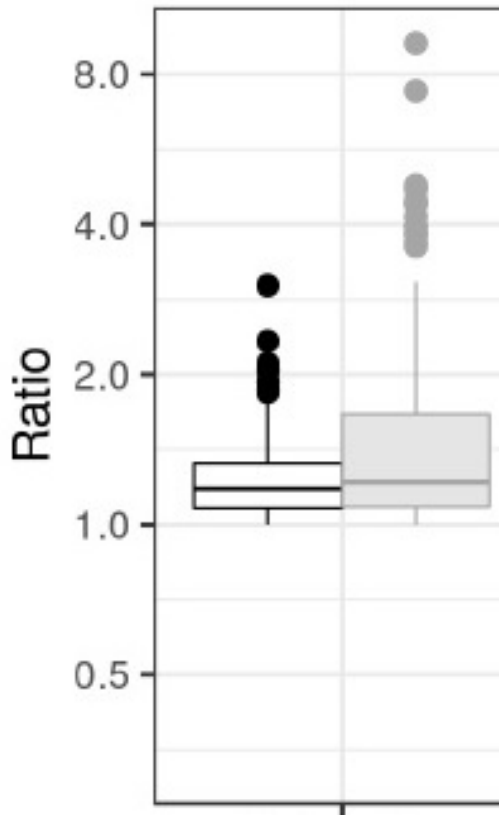


Estimation of relative AE risk: (1 - KM)

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.

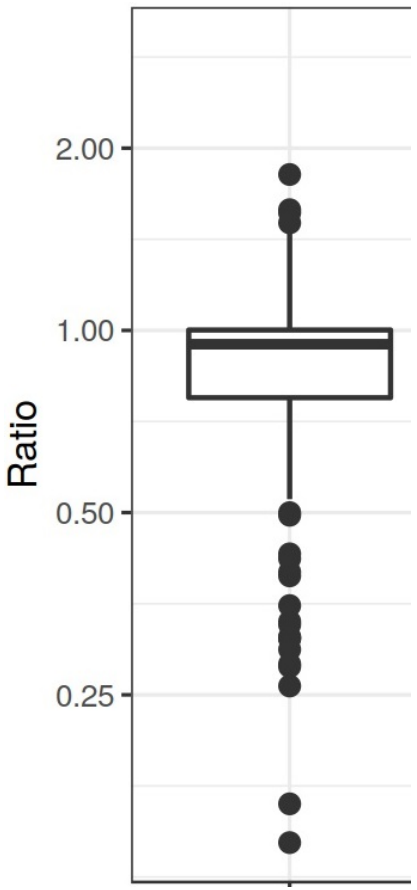


Estimation of relative AE risk: (1 - KM)

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.**

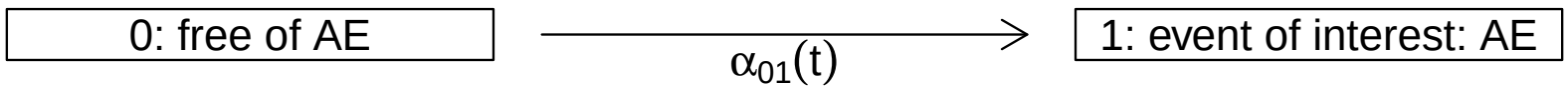


Now we have seen what does not work.

But what does work?

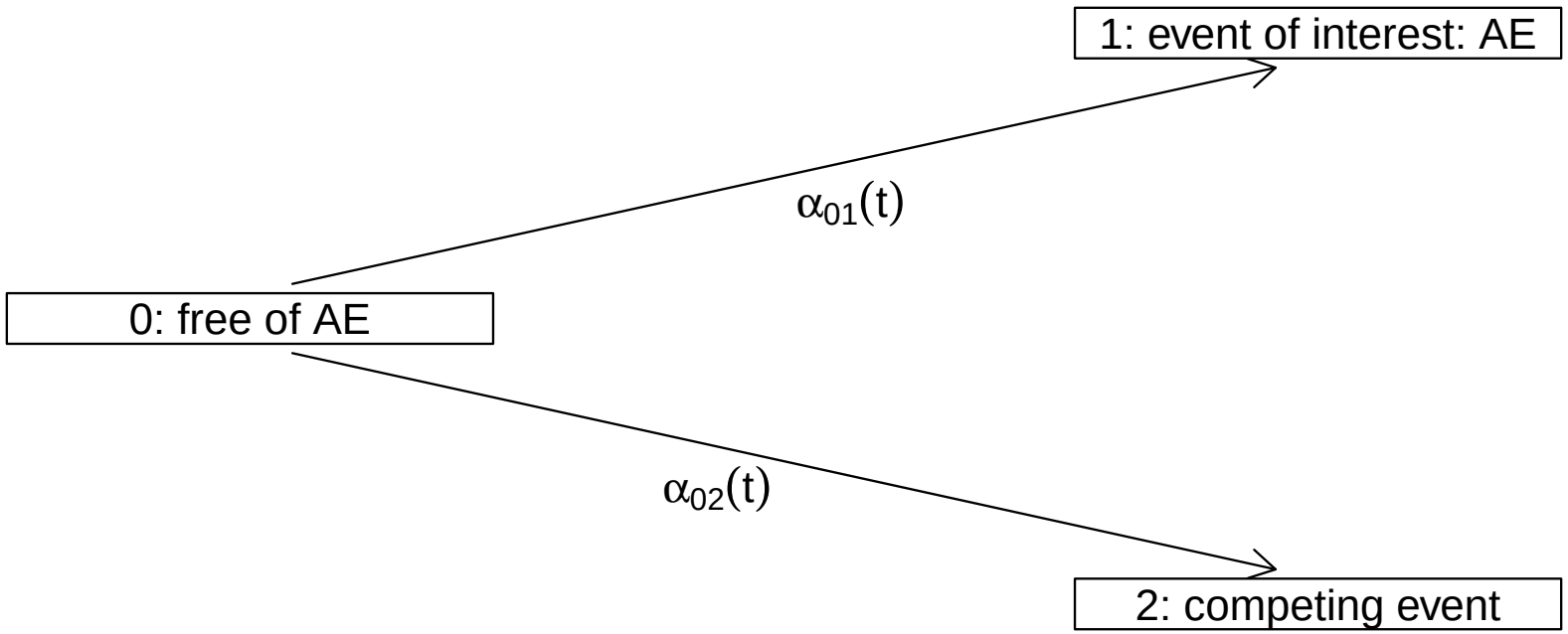
Canonical extension of survival analysis

Competing risks: generalize survival analysis from single combined endpoint to multiple first event types.



Canonical extension of survival analysis

Competing risks: generalize survival analysis from single combined endpoint to multiple first event types.



Doing now what patients need next

R version and packages used to generate these slides:

R version: R version 4.1.1 (2021-08-10)

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

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

This document was generated on 2022-05-04 at 17:00:15.